Corpus-Based Identification and Profiling of Vocabulary in Linguaskill Speaking Tests: A Methodological Framework for Adaptive Test Validation

Rosatiqah Yahya[1], Ng Yu Jin[1*], Mallika Vasugi A/P V. Govindarajoo[2], Kew Si Na[3] and Pipit Rahayu[4]

[1]UNITAR University College Kuala Lumpur, Malaysia

[2]UNITAR International University, Malaysia

[3]Universitiy Teknologi Malaysia, Malaysia

[4]Universitas Pasir Pengaraian, Indonesia

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100701181

Received: 26 July 2026; Accepted: 31 July 2026; Published: 24 August 2026

ABSTRACT

Computer-adaptive tests such as Cambridge Linguaskill adjust item difficulty dynamically to match test-taker proficiency, and their speaking components are now used for high-stakes admission and exit decisions, including in Malaysian higher education. Yet little research has examined the vocabulary load and lexical demands of adaptive speaking assessment, leaving open whether the vocabulary elicited at each reported level is congruent with the Common European Framework of Reference (CEFR) levels the scores claim to represent. This article develops a complete corpus-based methodological framework for identifying, classifying, and profiling the vocabulary of the Linguaskill Speaking Test. The framework specifies the compilation of a specialised corpus of transcribed candidate responses stratified across CEFR levels, supplemented by official test preparation materials, and a three-lens profiling procedure using LexTutor for BNC-COCA word-family analysis, the New General Service List for high-frequency lemma coverage, and Text Inspector for CEFR-aligned profiling against the English Vocabulary Profile. Operational criteria are defined for distinguishing core from peripheral vocabulary at each adaptive level, and decision rules are specified for classifying each level as aligned, over-demanding, or under-demanding relative to CEFR expectations, benchmarked against published lexical coverage thresholds for spoken English. The central hypothesis, motivated by prior coverage research, is that the lexical demands elicited by adaptive speaking tasks will not rise in step with claimed CEFR levels. The framework is grounded in an argument-based approach to validation and articulates the washback, fairness, and pedagogical stakes of the alignment question. The article contributes a replicable protocol for lexical validation of adaptive speaking tests and a benchmark synthesis for interpreting its future results.

Keywords: Linguaskill speaking vocabulary; LexTutor; Text Inspector; corpus-based approach; vocabulary threshold; CEFR alignment; adaptive assessment

INTRODUCTION

Language testing has moved decisively online, and with it has come the rise of computer-adaptive assessment. Tests such as Cambridge Linguaskill adjust the difficulty of what candidates encounter in real time, tailoring the measurement to the individual rather than administering a fixed form to all (Chapelle & Voss, 2016). The efficiency gains are considerable: shorter tests, faster results, and score reports mapped directly onto the levels of the Common European Framework of Reference for Languages (CEFR). These properties have carried adaptive tests into high-stakes use. In Malaysia, Linguaskill has been recognised by the Ministry of Higher Education for university admission purposes since 2020 and is used by institutions for admission benchmarking

and exit requirements alongside the Malaysian University English Test, a context in which the meaning of a reported CEFR level carries direct consequences for students' progression.

A reported CEFR level is a claim, and claims require validation. The argument-based approach to test validation holds that every link in the chain from performance to score to interpretation to use must be supported by evidence (Kane, 2013; Weir, 2005). For a speaking test, one link in that chain concerns vocabulary. Vocabulary knowledge is among the strongest predictors of second language speaking ability: receptive and productive vocabulary size predict oral proficiency and fluency (Koizumi & In'nami, 2013; Uchihara & Clenton, 2020; Uchihara & Saito, 2019), command of formulaic sequences distinguishes speakers judged fluent and proficient (Boers et al., 2006; Hougham et al., 2024), and lexical measures figure centrally in automated approaches to assessing spoken sophistication (Kyle & Crossley, 2015). If scores on a speaking test are to mean what CEFR levels say they mean, the vocabulary the test elicits and rewards at each level should be congruent with the lexical expectations of that level.

Whether this congruence holds for adaptive speaking tests is, at present, unknown. The lexical demands of receptive test input have been profiled in various contexts, and a substantial literature has established coverage benchmarks for spoken English, indicating that everyday conversation can be followed with knowledge of the most frequent 2,000 to 3,000 word families while comfortable comprehension requires 6,000 to 7,000 (Adolphs & Schmitt, 2003; Nation, 2006; van Zeeland & Schmitt, 2013; Webb & Rodgers, 2009a, 2009b). But the speaking components of adaptive tests have escaped comparable scrutiny.

Adaptivity itself complicates the question: because candidates at different levels encounter different tasks, the lexical demand of the test is not a single fixed property but a profile distributed across levels, and nothing guarantees that this profile ascends in step with the CEFR labels attached to the score bands. The abstract possibility that it does not, that adaptive mechanisms in speaking tasks do not require vocabulary demands congruent with claimed CEFR levels, is the central hypothesis motivating this research programme.

The stakes extend beyond psychometrics. Washback, the influence of a test on teaching and learning, presupposes that preparation aligned to the test is preparation aligned to the construct (Alderson & Wall, 1993); if the vocabulary rewarded by the test diverges from the vocabulary implied by its CEFR claims, preparation courses inherit the divergence.

Fairness is implicated because adaptive delivery must offer lexically equivalent demand to candidates routed through different task sequences. And pedagogy is implicated because teachers and materials writers in contexts such as Malaysia, where vocabulary limitations are a documented constraint on learners' English (Misbah et al., 2017; Tan & Goh, 2020; Wong et al., 2019), need to know which vocabulary genuinely matters for the assessments their students face.

This article makes the methodological contribution that such an investigation requires. It develops, in full and replicable detail, a corpus-based framework for identifying, classifying, and analysing the vocabulary of the Linguaskill Speaking Test. Because the empirical corpus is under compilation, the article is explicitly a framework and protocol contribution: it specifies the corpus design, the instruments, the analytical pipeline, the operational definitions, and the decision rules through which the alignment question will be answered, and it synthesises the published benchmarks against which the results will be interpreted. No empirical findings are claimed in advance of the data. Three research questions structure the conceptual paper are as follows:

- RQ1. What vocabulary is used in Linguaskill Speaking Test performance across CEFR levels, in terms of lexical types, frequency profiles, and CEFR-aligned classifications?
- RQ2. Which vocabulary items constitute the core, and which the periphery, of each adaptive level, and what lexical thresholds characterise each level?
- RQ3. To what extent do the lexical demands observed at each level align with the vocabulary expectations of the corresponding CEFR proficiency descriptors?

LITERATURE REVIEW

Vocabulary Knowledge and Second Language Speaking

The dependence of speaking ability on vocabulary is among the better replicated findings in second language research. Uchihara and Clenton (2020) found that productive vocabulary size predicted the speaking performance of university learners, and Uchihara and Saito (2019) showed that productive vocabulary knowledge related significantly to perceived oral ability, including comprehensibility. Koizumi and In'nami (2013) demonstrated the predictive role of vocabulary knowledge across fluency, accuracy, and syntactic complexity among novice to intermediate learners. Beyond size, the organisation of vocabulary matters: speech produced under real-time pressure leans heavily on formulaic sequences, and learners who deploy such sequences are judged more proficient and more fluent (Boers et al., 2006), with recent evidence indicating that lexical bundles contribute distinctively to oral fluency (Hougham et al., 2024).

Fluency itself is multi-componential, and temporal measures of speech relate systematically to judged proficiency (de Jong et al., 2012; Kormos & Dénes, 2004), while at the receptive interface, phonological vocabulary knowledge underpins listening proficiency (Saito et al., 2025; Stæhr, 2009). Lexical indices are correspondingly central to computational approaches to spoken proficiency measurement (Kyle & Crossley, 2015). For test validation, the implication is direct: the vocabulary elicited by speaking tasks is not incidental colouring but constitutive of the construct the tasks measure.

Lexical Coverage and Thresholds for Spoken English

How much vocabulary spoken English requires has been quantified in a coverage literature that parallels the better-known findings for reading. For written text, adequate comprehension is associated with 95 percent coverage of running words as a minimum and 98 percent as the level for comfortable independent processing (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011), figures that recent replication has qualified in detail but not overturned (Kremmel et al., 2023). For speech, the demands are lower but still substantial. Adolphs and Schmitt (2003) profiled conversational corpora and found that the most frequent 2,000 word families supply roughly 95 percent of running words in everyday spoken discourse, with around 3,000 needed to push coverage meaningfully higher. Nation (2006) estimated 6,000 to 7,000 word families for 98 percent coverage of spoken English, against 8,000 to 9,000 for written text.

van Zeeland and Schmitt (2013) connected these figures to comprehension experimentally, finding that listeners achieved adequate comprehension at 95 percent coverage, reachable with 2,000 to 3,000 families, while comfortable comprehension tracked the 98 percent level. Profiling of audiovisual input converges: 3,000 word families plus proper nouns yield approximately 95 percent coverage of television programmes and movies (Webb & Rodgers, 2009a, 2009b), and specialised inventories such as the academic spoken word list capture the additional demand of academic speech (Dang et al., 2017). Coverage profiles are also register-sensitive and variety-sensitive, shifting across text types within the same medium (Ha, 2022).

These benchmarks matter for the present framework in two ways. First, they supply the external standard against which the demand of test-elicited speech can be judged: a speaking task whose successful performance requires vocabulary well beyond the 2,000 to 3,000 family range is demanding more than everyday spoken English does. Second, they calibrate expectations across CEFR levels, since the vocabulary knowledge attributed to A2, B1, and B2 learners maps onto very different regions of the frequency scale, a mapping the English Vocabulary Profile makes explicit at item level.

The CEFR, Vocabulary Profiling, and Profiling Instruments

The CEFR describes proficiency at six levels through can-do descriptors, deliberately abstaining from language-specific inventories (Council of Europe, 2020). For English, that gap has been filled by two complementary enterprises. The English Profile programme identifies criterial features, the lexical and grammatical properties that distinguish adjacent levels of learner English (Hawkins & Filipović, 2012), and its English Vocabulary Profile (EVP) assigns CEFR levels to individual words and senses on the basis of learner corpus evidence,

providing the operational bridge between vocabulary and the framework (North, 2014). Web-based profiling instruments make these resources analytically usable. Text Inspector profiles texts against the EVP, returning the distribution of tokens across CEFR levels, and has been employed in published assessment research (Bax et al., 2019).

The Compleat Lexical Tutor (LexTutor) implements frequency-based profiling against the BNC-COCA word family lists, the standard instrument for coverage analysis (Cobb, 2007; Nation, 2017), while the New General Service List offers a modernised high-frequency lemma inventory with documented coverage advantages over its predecessor (Browne, 2014; Browne et al., 2013). Because word families, lemmas, and EVP senses are different counting units drawn from different reference corpora, profiles produced by the three instruments are complementary rather than redundant, and their joint use permits convergent validation of any lexical claim (Nation, 2013; Schmitt & Schmitt, 2014). Underlying all three is the multidimensional character of word knowledge itself, in which form, meaning, and use develop in describable relation to one another (González-Fernández & Schmitt, 2020), and the corresponding requirement that vocabulary measures be matched to the knowledge aspect at stake (Laufer & Goldstein, 2004; Laufer & Nation, 1999; Read, 2000).

Adaptive Language Assessment and Linguaskill

Computer-adaptive testing selects items conditional on the candidate's running performance, concentrating measurement where it is most informative. Two decades of technology-mediated assessment research have documented both the psychometric attractions of this design and its validation burdens, which include demonstrating that adaptively assembled test experiences support the same score interpretations as fixed forms (Chapelle & Voss, 2016). Vocabulary assessment itself contributed early demonstrations of computer-adaptive logic (Laufer & Goldstein, 2004), and candidate attitudes toward automated speaking assessment have been studied since the first generation of such instruments (Fan, 2014).

Linguaskill, launched globally by Cambridge in 2018 as the successor to BULATS, is a modular online multilevel test reporting on the Cambridge English Scale with CEFR-aligned bands. Its Speaking module lasts approximately 15 minutes across five task types and is scored through a hybrid model combining automated marking with examiner verification. Cambridge has published a validity argument for the Speaking test documenting its design, marking model, and evidence base (Cambridge University Press & Assessment, 2020). What that argument, and the wider literature, does not yet contain is a corpus-based lexical profile of the speech the test elicits across its levels. Given the centrality of vocabulary to the speaking construct established in Section 2.1, this is a substantive gap in the validation chain, and it is the gap the present framework is designed to close.

Corpus-Based Approaches to Test Validation

Corpus methods entered language assessment as instruments for describing what tests actually contain and elicit, complementing the judgement-based methods of traditional content validation. Specialised corpora of test input have been used to characterise the lexical demands of reading and listening materials, applying the same coverage logic later extended to television, movies, and news registers (Ha, 2022; Webb & Rodgers, 2009a, 2009b), and learner corpora underpin the level assignments of the English Profile programme itself (Hawkins & Filipović, 2012). Performance corpora, built from what candidates produce under test conditions, are rarer and more valuable: they capture the construct as realised rather than as intended, and they permit exactly the demand-side questions this study asks. The methodological requirements they impose are correspondingly stricter, since spoken performance data must pass through transcription, and profiling conventions must be fixed before analysis if results are to be replicable (Cobb, 2007; Nation, 2013).

Within this tradition, three design principles are settled and are adopted here. First, stratification must follow the interpretive units of the test, which for a CEFR-reporting instrument means level subcorpora. Second, register heterogeneity within a test must be preserved analytically rather than pooled away, since tasks as different as reading aloud and free monologue impose different lexical regimes. Third, external benchmarks must be fixed in advance, because a profile is meaningful only against a standard, and post hoc benchmark selection invites confirmation bias. Each principle is visible in the design choices of Section 5, and together they distinguish a validation-grade corpus study from an opportunistic word count.

The Research Gap

Three strands of the review converge on the gap. The coverage literature supplies benchmarks for spoken English but has not been applied to adaptive speaking test output. The CEFR profiling infrastructure supplies level-referenced lexical classification but has been applied principally to written and receptive materials. The validation literature demands evidence for each link from score to interpretation but contains no lexical-demand evidence for adaptive speaking components. No published study has compiled a corpus of Linguaskill Speaking performance, profiled it through frequency-based and CEFR-based lenses, and tested the congruence of its lexical demands with the levels its scores report. The framework developed below specifies exactly how that study is to be conducted.

CONCEPTUAL FRAMEWORK

Figure 1 presents the conceptual framework. Its spine runs from the adaptive assessment through the lexical demand it elicits to the CEFR meaning its scores claim, and the central construct is alignment: the degree to which elicited lexical demand rises in step with claimed level. The framework is an application of argument-based validation (Kane, 2013): the inference it interrogates is the one connecting observed speaking performance to a CEFR-levelled interpretation, and the evidence it generates is lexical. Three consequence domains hang from the alignment question. Washback validation asks whether preparation targeted at the test is thereby targeted at the construct (Alderson & Wall, 1993). Fairness asks whether the level claims are lexically equivalent for candidates routed through different adaptive experiences. Pedagogy asks what core vocabulary instruction should prioritise for learners facing the test, a question of immediate consequence in Malaysian tertiary contexts where Linguaskill is used for admission and exit decisions.

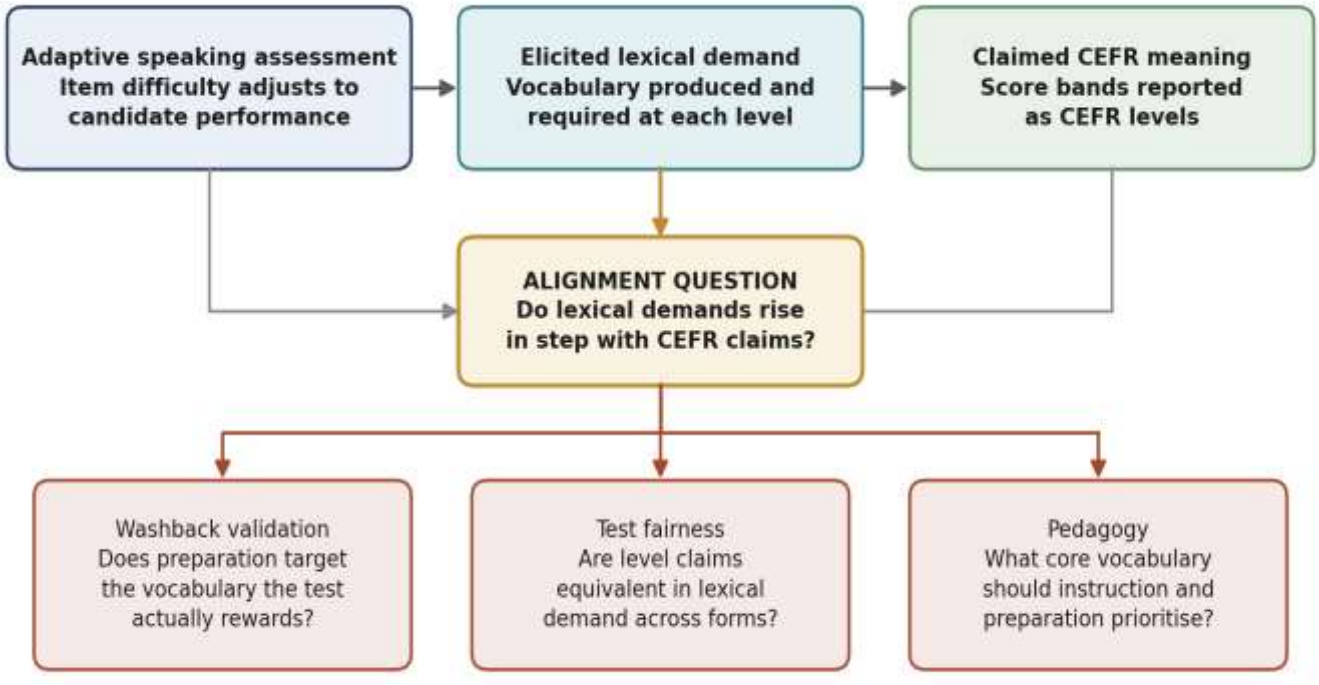

Figure 1: Conceptual framework linking adaptive assessment, lexical demand, CEFR alignment, and their consequence domains.

Two design obligations or suggested execution follow from the framework as above. First, because alignment is a relation between an observed profile and an expected one, the methodology must specify both sides: an empirical procedure for measuring elicited demand (Sections 5.2 to 5.5) and an explicit statement of level-referenced expectations drawn from the EVP and the coverage benchmarks (Section 6). Second, because misalignment can run in either direction, the classification scheme must distinguish over-demanding levels, whose vocabulary exceeds CEFR expectations, from under-demanding ones, whose vocabulary falls short, since

the two carry different consequences: over-demand threatens fairness to candidates at the claimed level, while under-demand inflates the meaning of the scores and misdirects washback. Locating this contribution requires acknowledging what a lexical lens does and does not see.

Speaking proficiency is multi-componential: alongside vocabulary, it comprises grammatical and syntactic complexity, the temporal dynamics of fluency, the organisation of discourse across a turn, pronunciation, and the communicative effectiveness with which a speaker meets a task (de Jong et al., 2012; Kormos & Denes, 2004). The present framework isolates the lexical component deliberately, not because the others are secondary, but because vocabulary is both a strong predictor of overall speaking ability (Koizumi & In'nami, 2013; Uchihara & Clenton, 2020) and the component for which a CEFR-referenced, corpus-based standard is most fully developed (Hawkins & Filipovic, 2012; North, 2014). Isolating one facet of a construct for rigorous measurement is a standard and defensible move, provided its partiality is stated plainly. The alignment verdict this framework produces is therefore a lexical verdict, one necessary input to a fuller validity argument rather than a substitute for it, and the design keeps that boundary explicit so that a favourable lexical alignment is never mistaken for global construct validity, and so that the framework can dock cleanly onto complementary analyses of the components it sets aside.

The Linguaskill Speaking Test as Object of Analysis

The Linguaskill Speaking module is delivered by computer in approximately 15 minutes and comprises five parts, summarised in Table 1 and Figure 2. Candidates answer interview questions about themselves, read sentences aloud, produce a one-minute long turn on a topic, describe visual or graphic material, and give opinions in a communication activity. Responses are recorded and scored through a hybrid model in which automated marking is verified by examiners (Cambridge University Press & Assessment, 2020). Results are reported on the Cambridge English Scale from 82 to 180 with a corresponding CEFR band, as shown in Table 2.

Table 1: Structure of the Linguaskill Speaking module.

Part	Task	What the candidate does	Speech elicited
1	Interview	Answers questions about themselves	Short personal-information turns
2	Reading aloud	Reads sentences aloud	Constrained oral reading
3	Long turn 1	Talks about a given topic for one minute	Extended monologue
4	Long turn 2	Describes a graphic or visual prompt	Extended monologue with visual support
5	Communication activity	Gives opinions on related questions	Opinion and argument turns

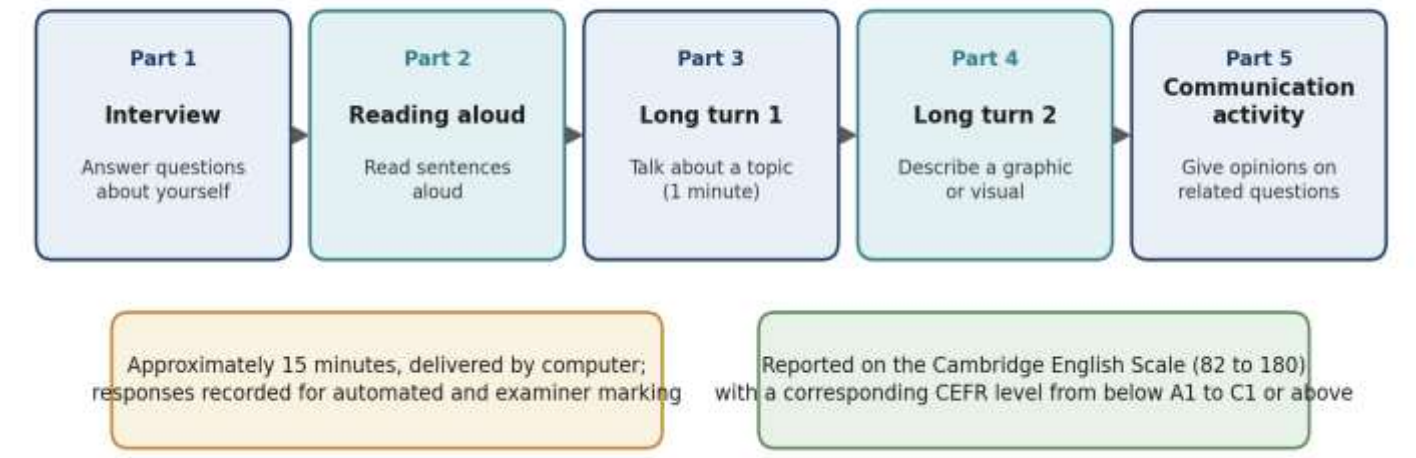

Figure 2: The five parts of the Linguaskill Speaking module and its reporting scale.

Table 2: Score bands and reported CEFR levels for Linguaskill.

Cambridge English Scale score	Reported CEFR level
180	C1 or above
160 to 179	B2
140 to 159	B1
120 to 139	A2
100 to 119	A1
82 to 99	Below A1

Two properties of the instrument shape the analytical design. First, the five parts elicit different registers of speech, from constrained oral reading to free monologue, so lexical profiles must be examinable both pooled and by part, since a demand difference concentrated in one task type is diagnostically different from a uniform shift. Second, because adaptivity distributes candidates across level-conditioned experiences, the corpus must be stratified by reported CEFR level, making the level subcorpus, rather than the individual response, the primary unit of profiling. These two requirements drive the sampling matrix specified in Section 5.2.

Adaptivity and the Distribution of Lexical Demand

Adaptivity changes what lexical demand means. In a fixed-form speaking test, the prompts are constant, and demand is a property of the form that can be profiled once. In an adaptive or multilevel instrument, candidates at different proficiency estimates encounter different prompts, so demand becomes a conditional property: the vocabulary the test requires of a B1 candidate is whatever the B1-conditioned tasks elicit, and the validation question is whether that conditional demand matches the B1 interpretation of the resulting score.

This is why the level subcorpus is the framework's unit of analysis, and why the input side matters alongside the output side: the reading-aloud sentences and prompt materials encountered at each level carry their own lexical profile, which the reference corpus of preparation materials approximates where the operational item bank cannot be accessed for security reasons.

The instrument's ongoing evolution reinforces the case for a dated, versioned baseline. Linguaskill has been revised since its 2018 launch, including changes introduced from 2024 that extend its certification uses, and any lexical profile is a profile of the version and item bank operative at the time of data collection. The framework therefore requires that the empirical report record the test version, collection window, and delivery mode alongside the corpus description, so that future replications can distinguish change in the test from change in the population taking it. Far from weakening the study, versioning makes it the first point in a monitoring series: the alignment classification of Table 7 can be re-run on each major revision, converting a one-off validation exercise into longitudinal quality assurance.

PROPOSED METHODOLOGY

Research Design

The study adopts a corpus-based descriptive and comparative design. It is descriptive in that it profiles the vocabulary of test-elicited speech through established quantitative instruments, and comparative in two directions: across the CEFR-levelled subcorpora, to establish whether lexical demand rises with level, and against external benchmarks, to establish whether the demand at each level matches CEFR expectations. Corpus methodology is appropriate because the research questions concern aggregate lexical properties of bodies of speech rather than the abilities of individual candidates, and because frequency- and profile-based measurement yields replicable results directly comparable with the published coverage literature (Cobb, 2007; Nation, 2013).

Corpus Construction

The primary corpus will consist of transcribed candidate responses to the Linguaskill Speaking Test, stratified by reported CEFR level from A2 to C1, the range within which adaptive speaking performance is concentrated. A supplementary reference corpus will be compiled from official Linguaskill test preparation materials, including published sample tasks and preparation guides, permitting comparison between the vocabulary candidates actually produce and the vocabulary preparation materials present. Table 3 specifies the sampling matrix. Within each level, responses will be sampled across all five test parts so that register variation is represented, and across candidates so that no individual dominates a subcorpus. Target subcorpus sizes are set at a minimum of 20,000 running words per level, a size at which frequency profiles of spoken data stabilise sufficiently for band-level comparison, with the pooled corpus targeted at approximately 100,000 running words, comparable to specialised spoken corpora used in published coverage research (Adolphs & Schmitt, 2003; Dang et al., 2017). Access to recorded responses will be arranged under institutional agreements, with all data anonymised at source.

Table 3: Proposed corpus sampling matrix.

Stratum	**CEFR level A2**	**CEFR level B1**	**CEFR level B2**	**CEFR level C1**	**Total**
Candidates sampled (minimum)	30	30	30	30	120
Test parts per candidate	5	5	5	5	
Target running words	20,000	25,000	27,500	27,500	100,000
Reference corpus (preparation materials)					20,000

Note. Level targets are minima; final sizes depend on response length, which itself increases with proficiency. The reference corpus is profiled separately and never pooled with performance data.

A corpus can license claims only about the population it represents, and the question of representativeness is therefore built into the sampling logic rather than deferred to the discussion. Three sources of heterogeneity are recorded as candidate-level metadata at the point of sampling: first language, prior formal instruction in English, and the educational programme through which each candidate entered the test. Recording these variables serves two ends at once. It allows the level subcorpora to be balanced so that no single first-language group or programme dominates a stratum, which guards against the risk that an apparent lexical signature of a level is in fact the signature of an over-represented subpopulation. It also turns a limitation into a research affordance, because the same metadata later permit a direct test of whether lexical performance at a given reported level varies systematically with candidate background, a question of immediate fairness relevance in multilingual settings such as Malaysia, where learners arrive through markedly different schooling trajectories (Misbah et al., 2017; Wong et al., 2019). Where the accessible pool proves narrower than the target, the empirical report will state the achieved distribution against the intended one, so that readers can judge the reach of every profile rather than infer a generality the data do not support.

Transcription and Cleaning Conventions

Spoken data enter lexical analysis only through transcription, and profiling results are sensitive to transcription decisions, so the conventions are fixed in advance and applied uniformly. Table 4 states them. Verbatim orthographic transcription is used, preserving the words candidates actually produce, including errors, since the object of analysis is elicited vocabulary rather than idealised speech. Hesitation markers and non-lexical fillers are tagged and excluded from profiling, following the treatment of marginal words in standard word lists (Nation, 2017). Proper nouns are retained but reported as a separate category, consistent with coverage research practice, because names are typically low-burden items inferable from context (Nation, 2013; Webb & Rodgers, 2009a). Unintelligible stretches are marked and excluded. Part 2 reading-aloud responses are transcribed but analysed separately from the free-production parts, since their vocabulary is supplied by the test rather than selected by the candidate, a distinction essential to interpreting demand.

Transcription is the point at which spoken performance becomes analysable text, and because every downstream profile inherits its errors, accuracy is treated here as a measurement property to be demonstrated rather than assumed. Beyond the double-transcription of a ten percent sample and the token-level agreement already specified, three further safeguards are adopted. Transcribers are trained against a shared conventions manual and a set of anchor recordings before production begins, so that early disagreements are resolved into rules rather than left to individual judgement. Ambiguous stretches, particularly those arising from accent, speaker overlap, or recording quality, are adjudicated by a second listener and, where they remain unresolved, are marked and excluded rather than guessed, since a plausible but incorrect word inflates or distorts the very frequency bands the study reports. A reliability check is then repeated at the midpoint of transcription, not only at its outset, because transcriber behaviour can drift over a long production window, and reporting agreement at more than one time point makes any such drift visible and correctable (Cobb, 2007; Nation, 2017).

Table 4: Transcription and cleaning conventions.

Convention	Rule	Rationale
Transcription mode	Verbatim orthographic; learner errors preserved	Object of analysis is elicited vocabulary as produced
Fillers and hesitations	Tagged (er, um, mm) and excluded from profiling	Marginal words distort frequency profiles
Proper nouns	Retained; reported as separate category	Low-burden items; standard coverage practice
Unintelligible speech	Marked as unclear; excluded	No defensible lexical assignment possible
Repetitions and false starts	Transcribed; duplicated tokens flagged for sensitivity analysis	Distinguishes lexical range from disfluency effects
Part 2 (reading aloud)	Transcribed; profiled separately from free production	Vocabulary is test-supplied, not candidate-selected
File format	Plain text, UTF-8, named by level and part	Tool compatibility and replicability

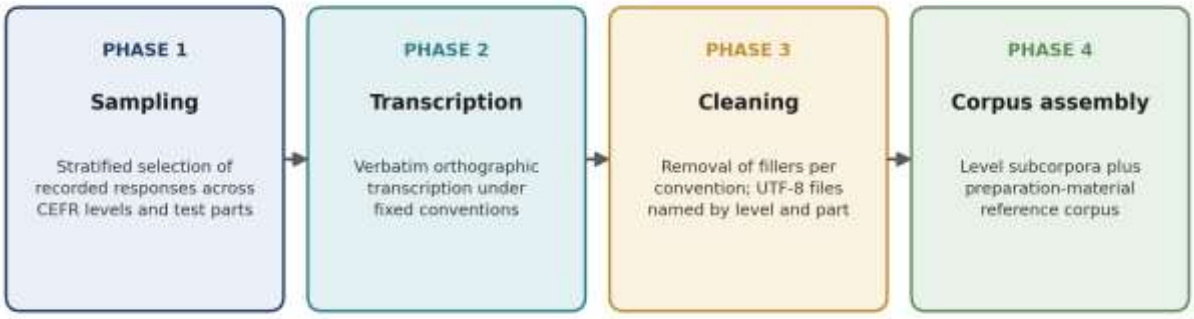

Figure 3: Corpus construction workflow.

Instruments and Measures

Three profiling lenses are applied to every subcorpus, summarised in Table 5. The first is frequency-based word-family profiling through LexTutor's VocabProfile against the BNC-COCA twenty-five band lists (Cobb, 2007; Nation, 2017), yielding the percentage of running words at each 1,000-family band, cumulative coverage curves, and the band at which each subcorpus reaches 95 and 98 percent coverage. The second is high-frequency lemma profiling against the New General Service List, yielding NGSL 1 to 3 and off-list shares (Browne, 2014; Browne et al., 2013). The third is CEFR-referenced profiling through Text Inspector against the English Vocabulary

Profile, yielding the distribution of tokens and types across EVP levels A1 to C2 (Bax et al., 2019). Alongside the three profiles, standard lexical statistics are computed for each subcorpus: tokens, types, lemmas, and length-corrected diversity, with type-token comparisons restricted to equal-sized samples to avoid the length dependence of diversity measures. Because the three lenses use different counting units and reference corpora, convergence across them constitutes strong evidence and divergence is itself diagnostic (Nation, 2013; Schmitt & Schmitt, 2014).

Table 5: Instruments and measures in the three-lens profiling procedure.

Lens	Instrument	Reference resource	Primary outputs
Word-family frequency	LexTutor VocabProfile	BNC-COCA 1,000-family bands (25 bands)	Band percentages; cumulative coverage; families needed for 95% and 98%
High-frequency lemmas	NGSL profiling	New General Service List (three 1,000-lemma bands)	NGSL 1 to 3 coverage; off-list share
CEFR classification	Text Inspector	English Vocabulary Profile (A1 to C2)	Token and type distribution across EVP levels; unlisted share
Lexical statistics	Corpus tools	Not applicable	Tokens, types, lemmas, length-controlled diversity

Analytical Procedure

Figure 4 summarises the pipeline. Each level subcorpus and the reference corpus pass through the three lenses. Two derived analyses follow. The first identifies core and peripheral vocabulary per level under the operational criteria of Table 6: an item is core to a level when it recurs across candidates and across free-production parts at that level, and peripheral when its occurrence is idiosyncratic. Dispersion across speakers, not raw frequency alone, is the defining property, since an item used often by one candidate is not evidence about the level. The second derived analysis compares demand across levels: cumulative coverage curves are plotted per level against the spoken-English benchmarks, and EVP profiles are compared for the expected rightward shift, that is, a growing share of B2 and above vocabulary as reported level rises.

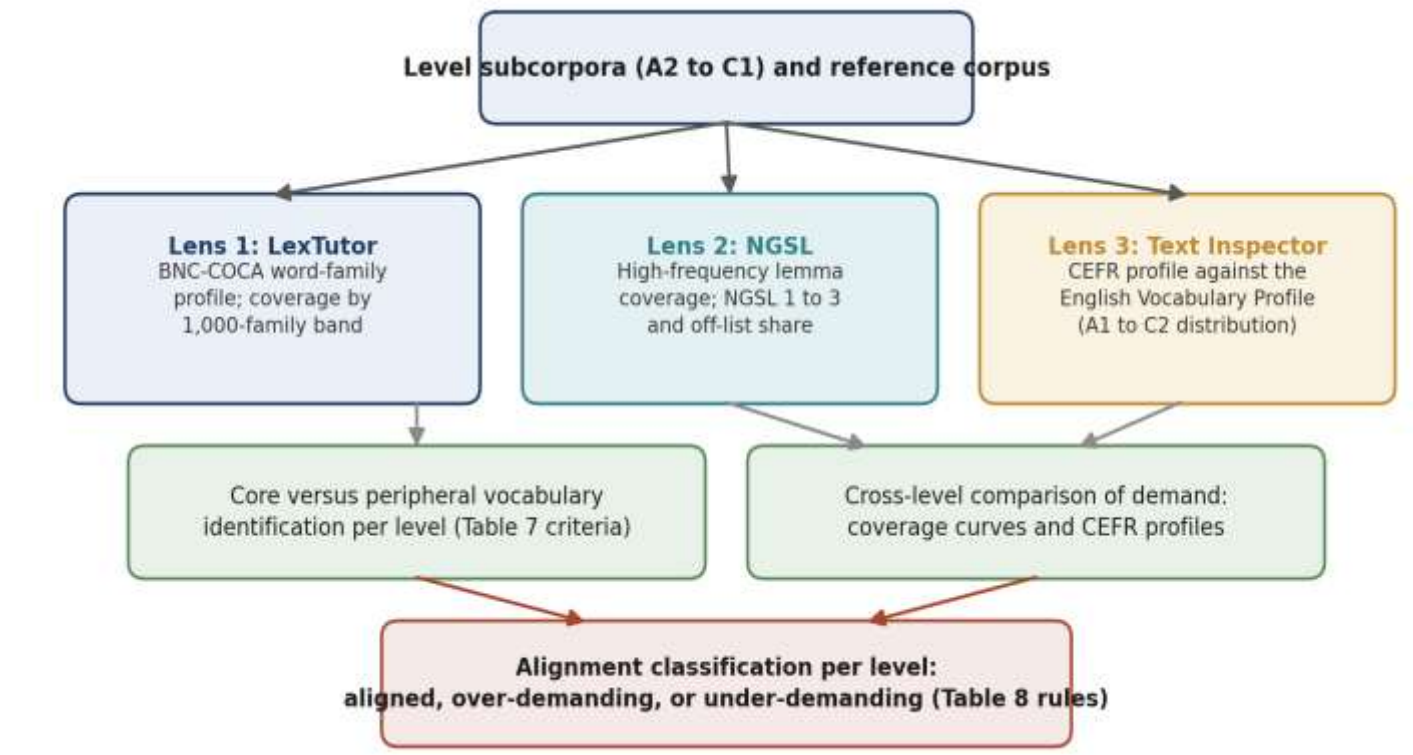

Figure 4: Analytical pipeline from level subcorpora to alignment classification.

Table 6: Operational definitions for core and peripheral vocabulary and level thresholds.

Category	Operational criterion	Interpretation
Core vocabulary of a level	Occurs in the speech of at least 30% of sampled candidates at the level and in at least two free-production parts	Vocabulary the level reliably elicits; the lexical signature of the level
Peripheral vocabulary of a level	Occurs below the dispersion criterion; concentrated in few candidates or a single part	Idiosyncratic or topic-bound items; not evidence of level demand
Test-supplied vocabulary	Items whose occurrence is confined to Part 2 reading aloud	Input demand of the test, analysed separately from production
Level threshold	Lowest BNC-COCA band at which the level subcorpus reaches 95% and 98% cumulative coverage	Vocabulary size characterising the level's demand

Alignment Decision Rules and Hypotheses

The alignment question requires explicit, pre-stated decision rules, given in Table 7. For each reported CEFR level, the observed EVP profile and coverage threshold are compared with the level's expectation: the modal EVP band of core vocabulary should sit at or immediately below the reported level, and the coverage threshold should fall within the range the spoken benchmarks associate with that proficiency region. A level is classified as aligned when both indicators fall within expectation, over-demanding when either indicator exceeds it, and under-demanding when either falls short. Applying the rules across levels yields the profile-level verdict on the central hypothesis. Three hypotheses, stated in Table 8, connect the analyses to the research questions; they are directional where the literature licenses direction and open where it does not.

Table 7: Decision rules for classifying each level's lexical alignment.

Indicator	Expectation for an aligned level	Over-demanding signal	Under-demanding signal
Modal EVP band of core vocabulary	At or immediately below the reported CEFR level	Modal band above reported level	Modal band two or more levels below
Coverage threshold (95%)	Within the benchmark range for the proficiency region (2,000 to 3,000 families at A2 to B1; rising toward mid-frequency at B2 to C1)	Threshold well above the region's range	Threshold flat across levels or below range
Cross-level trajectory	Monotonic rightward shift of EVP profile as level rises	Shift steeper than one CEFR band per level	No detectable shift between adjacent levels

Table 8: Hypotheses and their deciding analyses.

Hypothesis	Statement	Deciding analysis
H1	The lexical demands elicited by adaptive speaking tasks do not rise in step with claimed CEFR levels; adjacent levels will show substantially overlapping lexical profiles.	Cross-level comparison of EVP profiles and coverage curves (Table 7, trajectory rule)
H2	Core vocabulary at every level will be dominated by the first 2,000 word families, consistent with the coverage structure of spoken English generally.	Word-family profiles of core vocabulary per level against spoken benchmarks
H3	Preparation materials will present a lexical profile more demanding than candidate production at A2 to B1, generating a washback-relevant gap.	Reference corpus profile versus level subcorpora profiles

Ethical Considerations

Reliability is addressed at three points. Transcription reliability will be established by double-transcribing a 10 percent sample and reporting agreement at the token level, with conventions refined before full transcription proceeds. Profiling reliability is high by construction, since band assignment is algorithmic, but tool-version and list-version details will be recorded so that profiles are exactly reproducible (Cobb, 2007; Nation, 2017). Interpretive reliability is served by the pre-stated decision rules of Table 7, which remove post hoc discretion from the alignment classification. Ethically, the study analyses anonymised performance data under institutional data agreements; no candidate is identifiable, no new data collection from human participants occurs beyond the licensed use of recorded responses, and institutional ethical approval will be obtained before compilation begins. Test security is respected by reporting vocabulary at the level of profiles and dispersed items rather than reproducing test prompts.

Reporting Framework

To keep the eventual empirical report accountable to this protocol, its principal exhibits are specified now. The report will present, for each level subcorpus and the pooled corpus: the corpus description table (tokens, types, lemmas, candidates, parts represented); the BNC-COCA band distribution and cumulative coverage table with the 95 and 98 percent thresholds identified; the NGSL composition table; the EVP level distribution table; the core vocabulary inventory per level with dispersion statistics; the reference-corpus comparison; and the alignment classification per level with the Table 7 indicators shown explicitly. Figures will plot the cumulative coverage curves against the Figure 5 benchmarks and the EVP distributions across levels. Deviations from this reporting plan, should any prove necessary, will be declared and justified in the empirical article, preserving the auditability that pre-specification exists to provide.

Planned Pilot Demonstration

To narrow the distance between a protocol and its demonstrated effectiveness, the research programme includes a pilot phase in which the full pipeline is executed on a small authentic subset before the main corpus is compiled. The pilot draws a limited number of responses at two contrasting levels, for instance A2 and B2, and runs them through transcription, the three profiling lenses, and the alignment decision rules of Table 7 in exactly the form specified for the full study. Its purpose is not to answer the alignment question, which a small sample cannot settle, but to prove the operational chain end to end: to confirm that the transcription conventions survive contact with real recordings, that the three instruments return comparable and mergeable output on genuine candidate speech, and that the decision rules discriminate as intended when applied to data rather than to design. Reporting the pilot alongside the framework, with its inevitable frictions declared, gives readers evidence that the protocol is executable and not merely coherent on paper, and it allows the conventions to be refined once, cheaply, before the costly full compilation begins. This staged approach reflects the same pre-specification logic that governs the rest of the design (Kane, 2013): commitments are fixed in advance and then tested against reality in a low-stakes rehearsal, so that the eventual empirical article inherits a pipeline already shown to work.

Benchmarks for Interpretation

The framework's alignment judgements are only as good as the benchmarks they invoke, and this section fixes them in advance. Figure 5 and Table 9 synthesise the published coverage findings for spoken English that define the external standard. Everyday conversation reaches the 95 percent minimum with the most frequent 2,000 to 3,000 word families (Adolphs & Schmitt, 2003; van Zeeland & Schmitt, 2013); television and movies reach it with approximately 3,000 families plus proper nouns (Webb & Rodgers, 2009a, 2009b); and 98 percent coverage of spoken English requires 6,000 to 7,000 families (Nation, 2006; van Zeeland & Schmitt, 2013). Against this standard, speaking performance at A2 to B1 would be expected to draw overwhelmingly on the first 2,000 families, with B2 and C1 performance adding measurable mid-frequency and EVP upper-band material; a flat profile across levels would support H1, and demand exceeding the conversational benchmarks at lower levels would signal over-demand.

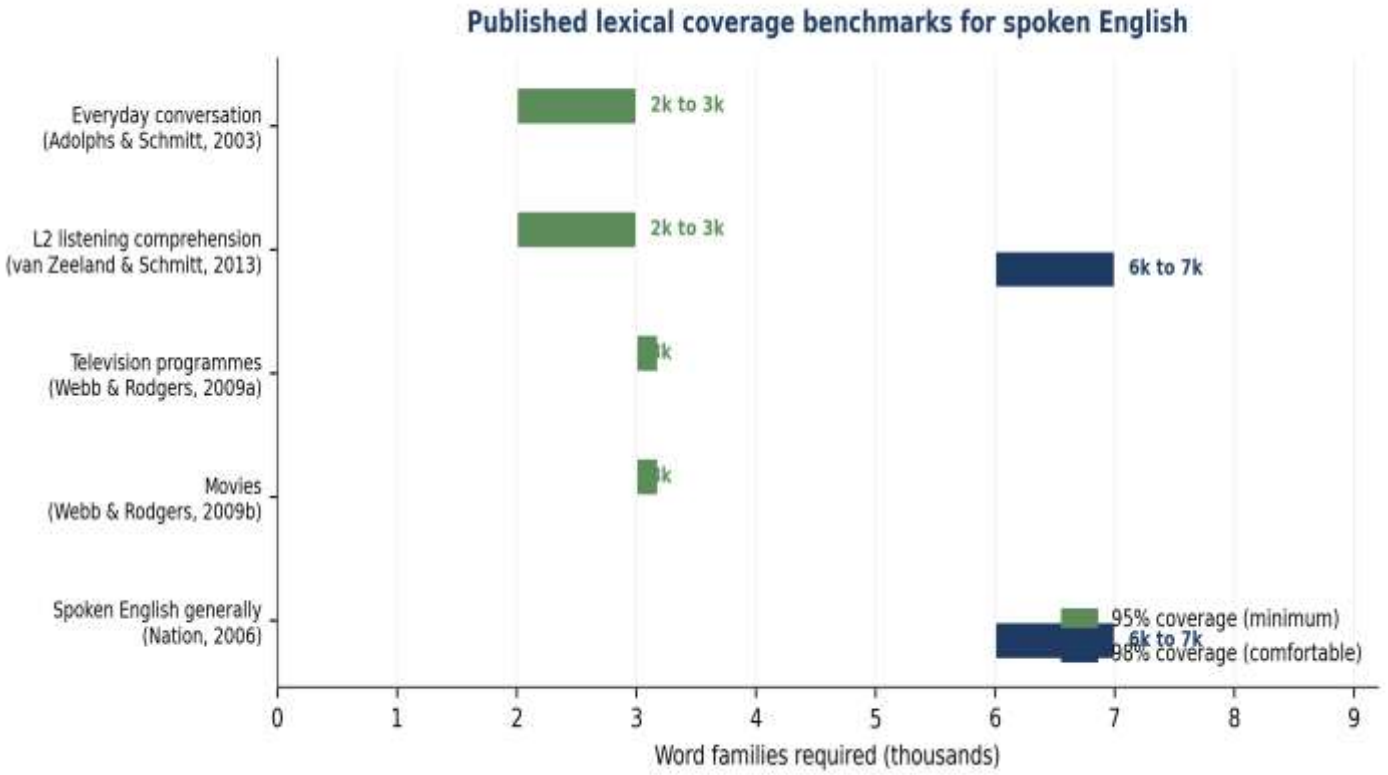

Figure 5: Published lexical coverage benchmarks for spoken English used as the interpretive standard.

Source	**Speech type**	**Families for 95% coverage**	**Families for 98% coverage**
Adolphs & Schmitt (2003)	Everyday conversation	About 2,000 to 3,000	Not estimated
van Zeeland & Schmitt (2013)	L2 listening comprehension	2,000 to 3,000	6,000 to 7,000
Webb & Rodgers (2009a)	Television programmes	3,000 plus proper nouns	Beyond mid-frequency
Webb & Rodgers (2009b)	Movies	3,000 plus proper nouns	Beyond mid-frequency
Nation (2006)	Spoken English generally	Within first 3,000 to 4,000	6,000 to 7,000

Table 9: Benchmark synthesis: vocabulary required for coverage of spoken English.

Note. Figures are as reported or derived in the cited studies; ranges reflect the corpora and counting conventions of each source. Proper nouns are treated as low-burden items throughout.

CEFR Level Expectations for Vocabulary Range

CEFR level	**Vocabulary range described by the framework**	**Operational expectation in this study**
A2	Sufficient vocabulary for routine, everyday transactions and basic personal needs	Core vocabulary overwhelmingly within the first 1,000 to 2,000 families; EVP profile centred on A1 to A2 items
B1	Sufficient vocabulary to express oneself, with some circumlocution, on most everyday topics	Core within the first 2,000 to 3,000 families; emerging B1 EVP share

B2	A good range of vocabulary for general topics and matters connected to one's field	Measurable mid-frequency component beyond 3,000 families; substantial B1 to B2 EVP share
C1	A broad lexical repertoire, including idiomatic expressions and colloquialisms	Visible presence beyond 5,000 families and of B2 and above EVP items, including formulaic material

Table 10: CEFR vocabulary-range expectations and their operationalisation.

Note. Range descriptions paraphrase the vocabulary range scale of the CEFR Companion Volume (Council of Europe, 2020); operational expectations combine those descriptors with the coverage benchmarks of Table 9.

Table 10 completes the expected side of the alignment relation. It translates the CEFR's qualitative vocabulary range descriptors into the quantitative terms of the profiling lenses, so that each level subcorpus can be compared against a stated expectation rather than an intuition. The translation is deliberately conservative: it claims only the orderings and regions that follow from the descriptors and the coverage literature jointly, and the alignment rules of Table 7 operate on those orderings rather than on any exact cut-point, which protects the classification from false precision.

A complementary expectation attaches to the EVP lens. Because the EVP assigns levels to word senses on learner corpus evidence, an aligned test should show the modal EVP band of each level's core vocabulary tracking the reported level, with A2 performance centred on A1 to A2 items and C1 performance showing a substantial B2 and above component (Hawkins & Filipović, 2012; North, 2014). The dual standard, frequency-based and CEFR-referenced, means that any alignment verdict rests on two independent measurement traditions, and the framework treats agreement between them as the criterion for a confident classification.

Implications

For Test Validation and Fairness

Whatever direction the results take, they feed the validity argument. Evidence of alignment would supply the missing lexical link in the chain from Linguaskill Speaking scores to CEFR interpretations, strengthening the argument published for the test (Cambridge University Press & Assessment, 2020) with independent corpus evidence of a kind argument-based validation explicitly calls for (Kane, 2013). Evidence of misalignment would localise the weakness: over-demand at particular levels raises fairness questions for candidates whose reported level presumes less vocabulary than the tasks in fact require, while under-demand would indicate that score meanings outrun elicited performance. Because the decision rules are pre-stated, either verdict is interpretable, and the same protocol applied to successive test versions would permit monitoring across the test's evolution.

For Pedagogy and Test Preparation

The core vocabulary inventories produced under Table 6 are directly teachable objects. For Malaysian tertiary programmes using Linguaskill for admission or exit decisions, level-specific core lists would let instructors target the vocabulary that performance at the next level reliably involves, replacing generic word lists with test-evidenced ones, and would ground that targeting in the instructed-vocabulary principles of repeated, evaluative engagement (Schmitt, 2008).

The comparison of candidate production with preparation materials under H3 speaks to washback directly: where preparation vocabulary and rewarded vocabulary diverge, materials can be recalibrated, converting washback from an assumed benefit into a validated one (Alderson & Wall, 1993). In a national context where vocabulary limitations demonstrably constrain learners (Misbah et al., 2017; Tan & Goh, 2020; Wong et al., 2019), the difference between preparing for the test and preparing for the construct is consequential, and this framework is designed to measure it.

For Research Methodology in Language Assessment

The framework also carries a methodological export. Its central design moves, stratification by reported level, register-preserving analysis, three-lens convergent profiling, pre-stated operational definitions, and pre-stated decision rules, constitute a template for lexical validation that transfers to any level-reporting speaking or writing assessment, adaptive or fixed-form. The insistence on fixing benchmarks and rules before data are seen imports the logic of pre-registration into corpus-based validation, where analytical flexibility is otherwise wide, and the explicit separation of test-supplied from candidate-selected vocabulary, enforced here through the treatment of the reading-aloud part, resolves an ambiguity that has been latent in test-corpus work generally. Publishing the protocol in advance of the results, as this article does, makes the subsequent empirical report auditable against its own commitments, a practice the validation literature endorses but corpus studies seldom implement (Kane, 2013).

The framework's value is not confined to a single instrument, and its robustness is best established comparatively. Applying the identical pipeline to other adaptive or multilevel speaking assessments would show whether any misalignment found in Linguaskill is a property of that test's item bank and marking model or a more general characteristic of how adaptive speaking tasks distribute lexical demand across levels. Because the design fixes its benchmarks, operational definitions, and decision rules before data are seen, it travels to a new instrument without renegotiation: only the corpus changes, while the interpretive standard against which each level is judged remains constant, which is precisely the condition under which cross-test findings become comparable rather than anecdotal (Chapelle & Voss, 2016; Kane, 2013).

Such comparison also carries a diagnostic payoff for the field. If several independent adaptive tests display the flat cross-level lexical profile that H1 anticipates, the finding would implicate the adaptive speaking paradigm itself rather than any one vendor, with consequences for how CEFR claims on such tests are warranted generally; if the tests diverge, the framework will have localised the difference to specific design choices that others can then examine. For research infrastructure, the framework identifies a reusable asset. A stratified, conventionally transcribed corpus of adaptive speaking performance has uses well beyond the alignment question: it supports formulaic-language analysis, disfluency research, automated scoring development, and cross-test comparison, and each subsequent use amortises the considerable cost of compliant compilation. Institutional partnerships of the kind this project requires are therefore best negotiated with multi-study reuse, under the original anonymisation and security constraints, written into the agreement from the outset.

Limitations and Future Research

The framework's limitations are stated with the same explicitness as its procedures. First, it is a protocol: until the corpus is compiled and profiled, the alignment question remains open, and this article claims only the framework, the benchmarks, and the hypotheses. Second, the corpus will represent the candidates whose responses are accessible under institutional agreements, and their first-language backgrounds and preparation histories will condition the vocabulary observed; the profile measures elicited demand for this population, not a universal property of the test. Third, transcription of learner speech involves irreducible judgement at the margins, which the double-transcription check quantifies but cannot eliminate. Fourth, the three profiling instruments inherit the limitations of their reference resources: BNC-COCA bands reflect general rather than test-specific frequency, the NGSL is a written-weighted inventory applied to speech, and the EVP assigns levels to senses on learner corpus evidence that is itself historical. Convergent use mitigates but does not abolish these inheritances. Fifth, lexical profiling captures single-item vocabulary more faithfully than formulaic language, whose contribution to speaking proficiency is documented (Boers et al., 2006; Hougham et al., 2024); a bundle-level extension is therefore a priority rather than an afterthought.

Future research follows in sequence. The immediate step is executing the protocol and reporting the profiles, thresholds, and alignment classifications with the corpus fully described. Extensions then include the formulaic-language analysis; replication across other adaptive tests to establish whether any misalignment is instrument-specific or endemic to adaptive speaking assessment; longitudinal application across Linguaskill versions; and an intervention strand testing whether instruction built on the empirically derived core lists improves both test performance and independently measured speaking proficiency, closing the loop from validation to pedagogy.

Because the corpus is transcribed and stratified once, it can be re-profiled for syntactic complexity and lexical sophistication using the automated indices developed for exactly this purpose (Kyle & Crossley, 2015), for fluency using the temporal measures whose relation to judged proficiency is well established (de Jong et al., 2012; Kormos & Denes, 2004), and for the formulaic and phonological dimensions whose contribution to oral ability is documented (Boers et al., 2006; Hougham et al., 2024; Saito et al., 2025).

Pronunciation and discourse coherence, which resist purely text-based analysis, would draw on the retained audio and on discourse-level annotation of the same responses. Read together, these strands would convert the single lexical verdict into a multi-dimensional profile of CEFR alignment, and the shared corpus ensures that each dimension is measured on the same performances rather than on separately assembled datasets, removing a confound that has limited earlier multi-construct comparisons. The lexical framework presented here is thus the first and most fully specified layer of a design intended from the outset to accommodate the others.

CONCLUSION

Adaptive speaking tests now carry high-stakes decisions on the strength of CEFR-labelled scores, yet the vocabulary those scores presuppose has never been profiled against the levels they claim. This article has built the complete methodological apparatus for doing so in the case of the Linguaskill Speaking Test: a stratified corpus design with fixed transcription conventions; a three-lens profiling procedure spanning BNC-COCA word families, NGSL lemmas, and the CEFR-referenced English Vocabulary Profile; operational definitions separating the core vocabulary of each level from its periphery; pre-stated decision rules classifying each level as aligned, over-demanding, or under-demanding; and a synthesis of the published coverage benchmarks for spoken English that anchors every judgement to the established literature.

The central hypothesis, that adaptive mechanisms in speaking tasks do not require vocabulary demands congruent with claimed CEFR levels, is now precisely testable, and the framework guarantees that whichever way the evidence falls, the answer will be interpretable, replicable, and consequential for washback validation, test fairness, and vocabulary pedagogy in the contexts, Malaysia prominent among them, where these scores shape students' academic futures. The contribution of this article is the instrument for that answer: a replicable protocol through which the lexical claims of adaptive speaking assessment can, for the first time, be held to account.

ACKNOWLEDGEMENT

The authors would like to thank UNITAR University College Kuala Lumpur for supporting this research.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship, or publication of this article.

REFERENCES

1. Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. Applied Linguistics, 24(4), 425–438. https://doi.org/10.1093/applin/24.4.425
2. Alderson, J. C., & Wall, D. (1993). Does washback exist? Applied Linguistics, 14(2), 115–129. https://doi.org/10.1093/applin/14.2.115
3. Bax, S., Waller, D., & Nakatsuhara, F. (2019). Researching L2 writers' use of metadiscourse markers at intermediate and advanced levels. System, 83, 79–95. https://doi.org/10.1016/j.system.2019.02.010
4. Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. Language Teaching Research, 10(3), 245–261. https://doi.org/10.1191/1362168806lr195oa
5. Browne, C. (2014). A new general service list: The better mousetrap we have been looking for? Vocabulary Learning and Instruction, 3(2), 1–10. https://doi.org/10.7820/vli.v03.2.browne

6. Browne, C., Culligan, B., & Phillips, J. (2013). The New General Service List. New General Service List Project. https://www.newgeneralservicelist.com
7. Cambridge University Press & Assessment. (2020). Linguaskill: Building a validity argument for the Speaking test. Cambridge Assessment English.
8. Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. Language Learning & Technology, 20(2), 116–128.
9. Cobb, T. (2007). Computing the vocabulary demands of L2 reading. Language Learning & Technology, 11(3), 38–63.
10. Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume. Council of Europe Publishing.
11. Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. Language Learning, 67(4), 959–997. https://doi.org/10.1111/lang.12253
12. de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. Studies in Second Language Acquisition, 34(1), 5–34. https://doi.org/10.1017/S0272263111000489
13. Fan, J. (2014). Chinese test takers' attitudes towards the Versant English Test: A mixed-methods approach. Language Testing in Asia, 4(1), Article 6. https://doi.org/10.1186/s40468-014-0006-9
14. González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. Applied Linguistics, 41(4), 481–505. https://doi.org/10.1093/applin/amy057
15. Ha, H. T. (2022). Lexical profile of newspapers revisited: A corpus-based analysis. Frontiers in Psychology, 13, Article 800983. https://doi.org/10.3389/fpsyg.2022.800983
16. Hawkins, J. A., & Filipović, L. (2012). Criterial features in L2 English: Specifying the reference levels of the Common European Framework. Cambridge University Press.
17. Hougham, D., Clenton, J., & Uchihara, T. (2024). Disentangling the contributions of shorter vs. longer lexical bundles to L2 oral fluency. System, 121, Article 103243. https://doi.org/10.1016/j.system.2024.103243
18. Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. Reading in a Foreign Language, 13(1), 403–430.
19. Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12000
20. Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. Journal of Language Teaching and Research, 4(5), 900–913. https://doi.org/10.4304/jltr.4.5.900-913
21. Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. System, 32(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001
22. Kremmel, B., Indrarathne, B., Kormos, J., & Suzuki, S. (2023). Unknown vocabulary density and reading comprehension: Replicating Hu and Nation (2000). Language Learning, 73(4), 1127–1163. https://doi.org/10.1111/lang.12622
23. Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Quarterly, 49(4), 757–786. https://doi.org/10.1002/tesq.194
24. Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. Language Learning, 54(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x
25. Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. Language Testing, 16(1), 33–51. https://doi.org/10.1177/026553229901600103
26. Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. Reading in a Foreign Language, 22(1), 15–30.
27. Misbah, N. H., Mohamad, M., Yunus, M. M., & Ya'acob, A. (2017). Identifying the factors contributing to students' difficulties in the English language learning. Creative Education, 8(13), 1999–2008. https://doi.org/10.4236/ce.2017.813136
28. Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? The Canadian Modern Language Review, 63(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59
29. Nation, I. S. P. (2013). Learning vocabulary in another language (2nd ed.). Cambridge University Press.
30. Nation, I. S. P. (2017). The BNC/COCA word family lists. Victoria University of Wellington.

31. North, B. (2014). The CEFR in practice. Cambridge University Press.
32. Read, J. (2000). Assessing vocabulary. Cambridge University Press.
33. Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2025). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. Studies in Second Language Acquisition, 47(1), 26–52. https://doi.org/10.1017/S0272263124000575
34. Schmitt, N. (2008). Instructed second language vocabulary learning. Language Teaching Research, 12(3), 329–363. https://doi.org/10.1177/1362168808089921
35. Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. Language Teaching, 47(4), 484–503. https://doi.org/10.1017/S0261444812000018
36. Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. The Modern Language Journal, 95(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x
37. Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. Studies in Second Language Acquisition, 31(4), 577–607. https://doi.org/10.1017/S0272263109990039
38. Tan, A. W. L., & Goh, L. H. (2020). Comparing the effectiveness of direct vocabulary instruction and incidental vocabulary learning in improving the academic vocabulary of Malaysian tertiary students. Pertanika Journal of Social Sciences and Humanities, 28(S2), 263–279.
39. Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. Language Teaching Research, 24(4), 540–556. https://doi.org/10.1177/1362168818799371
40. Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. The Language Learning Journal, 47(1), 64–75. https://doi.org/10.1080/09571736.2016.1191527
41. van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? Applied Linguistics, 34(4), 457–479. https://doi.org/10.1093/applin/ams074
42. Webb, S., & Rodgers, M. P. H. (2009a). Vocabulary demands of television programs. Language Learning, 59(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x
43. Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. Applied Linguistics, 30(3), 407–427. https://doi.org/10.1093/applin/amp010
44. Weir, C. J. (2005). Language testing and validation: An evidence-based approach. Palgrave Macmillan.
45. Wong, A. S. C., Lee, J. Y. V., Fung, M. E., & Willibrord, O. (2019). Vocabulary size of Malaysian secondary school students. Akademi Pengajian Bahasa, Universiti Teknologi MARA.
46. Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. Language Teaching Research, 26(4), 696–725. https://doi.org/10.1177/1362168820913998