Modeling CEFR Proficiency in a Speaking Test Corpus: Bayesian Ordinal Regression Analysis of Lexical, Syntactic, and Cohesion Indices

Fukuda Kohei†, Wei-tung Wang†, Naho Kawamoto‡, Yukio Tono†
†Tokyo University of Foreign Studies; ‡Chuo University
† 3-11-3, Asahi-cho, Fuchu, Tokyo 183-8514; ‡ 742-1, Higashi-Nakano, Hachioji, Tokyo 192-0393
† {fukuda.kohei.r0, winnie.wt, y.tono}@tufs.ac.jp; ‡ nkawamoto484@g.chuo-u.ac.jp

Abstract

This study examines which features of second language spoken English distinguish CEFR proficiency levels, and where on the scale they do so. The data are speaking-test responses from 700 test takers of BESTEP, a CEFR-referenced test developed by the Language Training and Testing Center in Taiwan and compiled into a corpus jointly with Tokyo University of Foreign Studies. Four automated tools computed indices of lexical sophistication (TAALES), lexical diversity (TAALED), cohesion (TAACO), and syntactic complexity and sophistication (TAASSC). Because CEFR levels form an ordinal scale, each index set was modelled with Bayesian ordinal regression, combining cumulative and adjacent-category models to estimate both overall effects and boundary-specific effects. With text length controlled, all four dimensions predicted CEFR level, but the lexical dimensions were among the strongest predictors and syntactic complexity the weakest. Boundary-specific analyses revealed a developmental pattern: function-word features and clause-level elaboration were more important at the lower boundary, whereas content-word features, source-text integration, and noun-phrase elaboration distinguished the higher levels. These level-dependent criterial features offer an empirical basis for validating and automatically scoring CEFR-referenced speaking tests.

Keywords: L2 speaking, speaking assessment, CEFR, criterial features, diversity, cohesion, complexity, ordinal regression

1. Introduction

The Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) has, in the two decades since its publication, become the de facto international standard for describing and comparing foreign language proficiency. Testing bodies around the world have carried out alignment studies linking their own scales to the six CEFR levels (Council of Europe, 2009), and these alignments now underpin score interpretation, admissions, and educational accountability in many jurisdictions. A recurring concern in this enterprise is the empirical basis of the levels themselves: for a CEFR-referenced test to be defensible, it is valuable to know which measurable features of learner language actually separate one level from the next — that is, to identify the *criterial features* that characterize each band (Hawkins & Filipović, 2012).

Taiwan offers a timely setting for this question. Under the government's Bilingual 2030 policy (National Development Council & Ministry of Education, 2021), which places particular emphasis on raising English proficiency in higher education, demand has grown for locally developed, CEFR-referenced assessments. To meet this need, the Language Training and Testing Center (LTTC) developed BESTEP (the BEST Test of English Proficiency), a four-skill test that targets university and graduate students and reports results on the CEFR scale, and BESTEP has since been adopted by a substantial number of universities across Taiwan.

The present study draws on the speaking component of BESTEP. Working jointly, LTTC and Tokyo University of Foreign Studies compiled a corpus of spoken English from 700 test takers whose responses had been rated on the CEFR. One aim of this collaboration is to examine the validity of the correspondence between test scores and CEFR levels by analyzing, level by level, the linguistic characteristics of the learners' speech, and thereby to identify CEFR-based criterial features that can serve as a foundation for the future automated scoring of the BESTEP speaking test.

As a first step towards this goal, the present study works from full transcriptions of the corpus and examines a set of representative linguistic indices that have been widely used in previous research on speaking proficiency. Because CEFR levels form an ordered rather than an interval scale, we treat CEFR level as an ordinal outcome and model it through Bayesian ordinal regression — an approach that estimates not only the overall contribution of each index but also where on the proficiency scale it discriminates between adjacent levels.

2. Literature Review

At the intersection of second language acquisition research and language assessment, natural language processing tools have made it possible to compute linguistic features automatically from learner production and to relate them to human ratings of proficiency. Tools are now available for lexical sophistication (TAALES; Kyle & Crossley, 2015; Kyle et al., 2018), lexical diversity (TAALED; Kyle et al., 2021), cohesion (TAACO; Crossley et al., 2016; Crossley et al., 2019), and syntactic complexity and sophistication (TAASSC; Kyle, 2016), which allows the linguistic basis of proficiency ratings to be examined with fine-grained indices.

This line of research has yielded a number of findings for each dimension. Lexical sophistication is a multidimensional construct that word frequency alone cannot capture: indices such as n-gram association strength predict proficiency more strongly than frequency (Kyle & Crossley, 2015; Kyle et al., 2018), and content-word and function-word indices separate into distinct dimensions, with function-word indices uniquely predicting proficiency in spoken production (Kim et al., 2018; Eguchi & Kyle, 2020; Eguchi, 2022). For lexical diversity, Lu (2012) and Bulté and Roothooft (2020) reported that the effect is particularly large in ratings of speaking. For cohesion, local indices such as connectives and lexical overlap show no relation, or a negative relation, to ratings on independent

tasks (Crossley & McNamara, 2013; Guo et al., 2013), whereas on source-based integrated tasks the degree of similarity to and integration of the source text is a positive predictor (Crossley et al., 2014; Crossley et al., 2019). For syntactic complexity, phrasal-level complexity predicts quality ratings better than clausal-level complexity in writing (Biber et al., 2011; Kyle & Crossley, 2018), whereas in speaking it discriminates proficiency less effectively than lexical and fluency measures (Iwashita et al., 2008).

Three issues nevertheless remain unresolved. First, most of these findings are based on written language, and they may not transfer directly to speech, as the reversal in the direction of frequency-based lexical sophistication effects illustrates (Eguchi & Kyle, 2020). Second, few studies have compared multiple dimensions simultaneously within the same dataset. Kolahi Ahari et al. (2025) is one of the few exceptions. This study found that lexical diversity and lexical sophistication were strong predictors of L2 speaking proficiency, that syntactic complexity had no direct effect, and that the relative importance of each construct varied with proficiency level. Third, few studies have asked which boundaries on the proficiency scale a given index discriminates. Eguchi (2022) is methodologically the closest to the present study in modeling CEFR-based ratings with Bayesian ordinal regression, but the estimation of boundary-specific effects was abandoned due to the sample size.

The present study therefore analyzes indices from all four dimensions within a single Bayesian ordinal regression framework, using speaking test responses rated on the CEFR (N = 700). The following research questions are addressed.

RQ1: Do indices of lexical sophistication, lexical diversity, cohesion, and syntactic complexity and sophistication predict CEFR speaking proficiency once text length is controlled for, and does their relative predictive power differ across the four dimensions?
RQ2: Within each dimension, which indices are associated with CEFR level?
RQ3: Do the discriminating effects of the indices differ across CEFR proficiency boundaries?

3. Method

3.1 Data

The data were drawn from the speaking test part of BESTEP (BEST Test of English Proficiency), which is developed and administered by the Language Training and Testing Center (LTTC) in Taiwan. BESTEP targets university and graduate students in Taiwan and measures English proficiency in line with CEFR Levels A through C. The speaking test comprises three parts and eight items: six short questions about daily life and learning (Part 1), one opinion-expression task on a topic related to daily life and learning based on charts and three written questions provided (Part 2), and one presentation task requiring the integration of information from a written text and from graphs and tables on a more abstract topic (Part 3). The test takes approximately 15 minutes, and total scores are converted into 10 CEFR levels ranging from Below A1 to C1.

Manually transcribed responses from 700 test takers were analyzed, drawn in equal numbers from two administrations with different topics (Forms S-2402 and S-2404, n = 350 each). Although the eight items are independent, all responses from a given test taker were concatenated and treated as a single text for automated linguistic analysis, which averaged 458.6 words in length (SD = 201.5). The distribution of CEFR levels was skewed, with a mode at B2 (n = 186, 26.6%) and few test takers at the lowest levels (Table 1). Because Below A1, A1, and A1+ together accounted for only 56 test takers, they were collapsed into a single level, so that CEFR level was treated as an eight-level ordinal factor with seven boundaries between adjacent levels.

CEFR level	*n* (%)	Analyzed level	Three-band analysis
Below A1	9 (1.3)	A1_or_below	excluded
A1	11 (1.6)	A1_or_below	excluded
A1+	36 (5.1)	A1_or_below	excluded
A2	95 (13.6)	A2	A2
A2+	110 (15.7)	A2+	A2
B1	65 (9.3)	B1	B1
B1+	79 (11.3)	B1+	B1
B2	186 (26.6)	B2	B2
B2+	83 (11.9)	B2+	B2
C1	26 (3.7)	C1	excluded
Total	**700**	8 levels	3 bands (618)

Table 1: Distribution of CEFR levels (N = 700) and the two collapsing schemes used in the analyses.

3.2 Linguistic Indices

Four tools, each corresponding to one construct, were applied to the transcripts: lexical sophistication (TAALES v2.8.1; Kyle & Crossley, 2015; Kyle et al., 2018), lexical diversity (TAALED v1.4.1; Kyle et al., 2021), cohesion (TAACO v2.1.3; Crossley et al., 2016; Crossley et al., 2019), and syntactic complexity and sophistication (TAASSC v1.3.8; Kyle, 2016). This pairing of tools with constructs follows Kolahi Ahari et al. (2025).

3.3 Index Selection

Each tool outputs a large number of indices, many of which are conceptually redundant variants that repeat the same measurement with a different reference corpus, unit of analysis, or part-of-speech category. Candidates were therefore narrowed in two stages, following the logic of Eguchi and Kyle (2020).

Candidates were first selected manually so as to cover the subconstructs of each tool while excluding redundant variants and indices inappropriate for transcribed speech. For TAALES, frequency and age-of-acquisition indices were retained in their separate content-word (CW) and function-word (FW) forms rather than in all-words form, because CW and FW indices have been shown to form independent dimensions in spoken production (Eguchi & Kyle, 2020; Eguchi, 2022); reference corpora were restricted to COCA spoken, COCA academic, and SUBTLEXus, given that the data come from a speaking test for academic purposes; and all n-gram association measures were retained, as none is established as superior. For TAALED, length-dependent indices such as simple

TTR were excluded in favor of length-robust indices (MATTR, HD-D, and MTLD). For TAACO, paragraph-level lexical overlap indices were excluded because paragraph boundaries are not clear in transcripts, and all connective indices were retained. For TAASSC, noun phrase indices were retained only at the level of grammatical function, such as dependents per nominal, rather than for individual dependency types, and verb-argument construction (VAC) indices only in their COCA academic version, as no spoken counterpart was available. This procedure yielded 54 candidates for TAALES, 100 for TAACO, 17 for TAALED, and 32 for TAASSC.

A two-step data-driven screening was then applied independently within each tool. In Step 1, indices whose Spearman correlation with CEFR ordinal values fell below a threshold were removed. In Step 2, Spearman correlations among the remaining indices were computed, and pairs whose correlation exceeded a threshold were resolved iteratively: at each iteration, the pair with the highest intercorrelation was identified, and of the two indices the one with the weaker correlation with CEFR was discarded. This process was repeated until no pair exceeded the threshold. Word count was included among the candidates for every tool. The standard thresholds were relaxed for two tools: for TAALED, correlations among indices were so high that a Step 2 threshold of .60 would have removed both focal HD-D indices, and for TAASSC, a Step 1 threshold of .20 removed too many indices at that stage. For TAACO, the 21 indices surviving Step 2 were judged too numerous, and an additional manual review of conceptual redundancy and of partial correlations controlling for word count reduced them to 12. Table 2 reports the thresholds and the number of indices retained at each step, and Table 3 lists the final sets.

Tool	Step 1	Step 2	Candidates	After Step 1	After Step 2
TAALES	< .20	> .60	54	46	17
TAACO	< .20	> .60	100	66	21 → 13 (reviewed)
TAALED	< .20	> .70	17	15	3
TAASSC	< .10	> .70	32	29	16

Table 2: Screening thresholds (absolute Spearman ρ) and number of indices retained.
Counts after Step 2 include word count.

Tool (construct)	Selected indices
TAALES (lexical sophistication ; 16)	*Academic language*: Core_AFL_Normed, Spoken_AFL_Normed. *Age of acquisition*: Kuperman_AoA_CW. *Contextual distinctiveness*: McD_CD. *N-gram association strength*: COCA_spoken_bi_MI, COCA_spoken_bi_DP, COCA_spoken_tri_MI2, COCA_spoken_tri_2_MI, COCA_spoken_tri_2_DP. *N-gram range*: COCA_Academic_Bigram_Range_Log, COCA_Academic_Trigram_Range_Log. *Psycholinguistic norms*: Brysbaert_Concreteness_Combined_CW, Brysbaert_Concreteness_Combined_FW. *Word frequency and range*: SUBTLEXus_Freq_FW_Log, SUBTLEXus_Range_FW_Log. *Word recognition norms*: WN_Mean_Accuracy
TAACO (cohesion; 12)	*Source text similarity*: source_similarity_word2vec, mag_news_uni_keywords_percentage. *Givenness*: unattended_demonstratives, determiners, repeated_content_and_pronoun_lemmas. *Connectives*: lexical_subordinators, all_logical, all_temporal, opposition. *Lexical overlap*: adjacent_overlap_2_verb_sent. *Semantic overlap*: syn_overlap_sent_noun, lsa_2_all_sent
TAALED (lexical diversity; 2)	*Lexical diversity*: hdd42_cw, hdd42_fw
TAASSC (syntactic complexity and sophistication ; 15)	*Unit length*: MLC. *Clausal and phrasal (SCA)*: T_S, CP_T, CN_C. *Clause complexity and variety*: cl_av_deps, cl_ndeps_std_dev. *Noun phrase complexity and variety*: av_dobj_deps, nominal_deps_stdev, nsubj_stdev, dobj_stdev. *Syntactic sophistication*: acad_av_construction_freq_log, acad_av_lemma_construction_freq_log, acad_lemma_construction_attested, acad_av_delta_p_verb_cue_stdev, acad_av_delta_p_const_cue_stdev

Table 3: Focal predictors retained for each tool, grouped by index category.

Because both screening steps were carried out on the full dataset and relied on each index's rank correlation with CEFR to decide retention, the predictor sets were selected with reference to the same outcome that the models below predict. The focal effect sizes reported in Section 4.2 and the model comparisons in Section 4.4 should therefore be read as potentially optimistic, a point to which we return in Section 5.5.

3.4 Statistical Analysis

CEFR levels are ordered categories, and the distances between adjacent levels cannot be assumed to be equal. Liddell and Kruschke (2018) showed that analyzing ordinal data with metric models systematically produces false alarms, misses, and even reversals in which the estimated direction of an effect is opposite to the true direction, and that such errors cannot be identified from the output of a metric model alone. Bürkner and Vuorre (2019) similarly argue that an ordinal outcome is better modeled as a continuous latent variable partitioned by category-specific thresholds. CEFR level was accordingly modeled as an ordinal rather than a metric outcome.

Two families of Bayesian ordinal regression model were used in a complementary fashion. The cumulative model assumes a continuous latent variable partitioned into *K* ordinal categories by K-1 thresholds and, with a logit link, estimates the change in the log cumulative odds of falling in a higher category; because it imposes the proportional odds assumption, it captures the overall direction and magnitude of each effect but not effects that vary across boundaries. The adjacent-category model instead models the log odds of the ratio of selection probabilities between two adjacent categories, and the cs() function of brms allows each predictor a separate slope at each boundary. Category-specific effects are not available in the cumulative model, so boundary-specific effects were estimated with the adjacent-category model.

The same procedure was used for all four tools; only the set of focal predictors differed. Each main model contained that tool's selected indices as focal predictors, word count as a control, and test form as a fixed-effect control; all

numeric predictors were *z* standardized. Text length was expected to have the largest effect but is not itself of interest here. For each tool and family, this main model was compared with a null model containing the control variables only. Weakly informative normal(0, 1) priors were placed on the slopes for TAALES, TAALED, and TAASSC; for TAACO this prior yielded prior predictive distributions skewed toward the extreme CEFR categories, so it was tightened to normal(0, 0.5). Thresholds received a student_t(3, 0, 2.5) prior throughout. Models were estimated in Stan via brms (Bürkner, 2017) with four chains of 2,000 iterations each (1,000 warmup), giving 4,000 posterior draws, with adapt_delta set to .95 for cumulative and .99 for adjacent-category models and max_treedepth set to 12.

All models were fitted to the eight-level outcome (N = 700), and every model comparison reported below is based on these eight-level models. In the adjacent-category models, however, the sparse categories at the extremes left several boundary estimates unstable. Boundary-specific effects were therefore re-estimated after excluding A1_or_below and C1 and collapsing the remaining six levels into three bands, A2 (A2/A2+), B1 (B1/B1+), and B2 (B2/B2+), which leaves two boundaries with adequate cell counts (n = 618; Table 1). This gains stability at the cost of restricting the range to A2–B2+ and of losing discrimination within a band.

Convergence was assessed against three criteria: R-hat below 1.01 for all parameters, sufficient bulk and tail effective sample sizes (ideally above 1,000 and at minimum above 400), and good mixing in trace plots. Posterior predictive checks were run for every model. Models were compared by PSIS-LOO cross-validation (Vehtari et al., 2017) with loo_compare(), and, following Bürkner and Vuorre (2019), a difference in predictive accuracy was treated as clear when the absolute value of elpd_diff was at least about twice its standard error. Comparisons were made along three axes: null against main model within each tool and family, across tools within a family, and across families within a tool. All compared models were fitted to the same 700 responses and the same outcome.

4. Results

4.1 Model Diagnostics

All models converged: R-hat was below 1.01 for every parameter, bulk and tail effective sample sizes exceeded 1,400 throughout (minimum 1,570 across the cumulative models), no divergent transitions occurred, and trace plots showed good mixing across chains. Posterior predictive checks reproduced the observed category frequencies closely in both model families, including when stratified by test form. The posterior summaries below are therefore stable.

4.2 Overall Effects (Cumulative Model)

Text length was the strongest positive predictor in every model (OR = 12.94–19.11), so the focal effects reported below are all estimated with text length held constant. Test form was credible for TAALES (b = 0.56, 95% CrI [0.11, 1.02]) and TAASSC (b = −0.37, [−0.70, −0.05]) but not for TAACO or TAALED. Table 4 reports the focal predictors whose CrIs excluded zero.

For TAALES, 4 of the 16 lexical sophistication indices were credible. Bigram association strength in spoken COCA (COCA spoken bi MI) had the strongest positive effect (OR = 2.38), followed by the age of acquisition of content words (Kuperman AoA CW; OR = 1.75). Function-word concreteness (Brysbaert Concreteness Combined FW) was negative (OR = 0.65), so that more abstract function-word vocabulary was associated with higher CEFR levels, whereas the corresponding content-word index was not credible. Function-word dispersion (SUBTLEXus Range FW Log) was weakly negative (OR = 0.82). The remaining 12 indices, including the academic formula indices, the other n-gram indices, and word naming accuracy, were not credible.

For TAACO, 4 of the 12 cohesion indices were credible. Semantic similarity to the source text (source similarity word2vec) showed the strongest effect (OR = 1.99): responses reflecting more of the content of the texts supplied in the integrated tasks were associated with higher CEFR levels. Subordinating conjunctions (lexical subordinators; OR = 1.56), contrastive and adversative connectives (opposition; OR = 1.34), and pronominal demonstratives (unattended demonstratives; OR = 1.26) were also positive.

For TAALED, both lexical diversity indices were credible and positive, and the diversity of function words (hdd42 fw; OR = 3.81) had a larger effect than that of content words (hdd42 cw; OR = 2.04).

For TAASSC, 5 of the 15 syntactic indices were credible. Complex nominals per clause (CN_C; OR = 1.75) had the strongest positive effect, followed by dependents per clause (cl av deps; OR = 1.44), the attestation rate of verb lemma–VAC combinations in the reference corpus (acad lemma construction attested; OR = 1.24), and the variability of dependents in subject noun phrases (nsubj stdev; OR = 1.20). Dependents per direct object (av dobj deps; OR = 0.81) was negative.

Tool	Predictor	b	95% CrI	OR
TAALES	COCA spoken bi MI	0.87	[0.63, 1.11]	2.38
TAALES	Kuperman AoA CW	0.56	[0.35, 0.77]	1.75
TAALES	SUBTLEXus Range FW Log	−0.20	[−0.38, −0.03]	0.82
TAALES	Brysbaert Concreteness Combined FW	−0.43	[−0.63, −0.22]	0.65
TAACO	Source similarity word2vec	0.69	[0.51, 0.88]	1.99
TAACO	Lexical subordinators	0.45	[0.25, 0.64]	1.56
TAACO	opposition	0.29	[0.14, 0.45]	1.34

Tool	Predictor	b	95% CrI	OR
TAACO	Unattended demonstratives	0.23	[0.05, 0.41]	1.26
TAALED	hdd42 fw	1.34	[1.01, 1.67]	3.81
TAALED	hdd42 cw	0.71	[0.39, 1.08]	2.04
TAASSC	CN C	0.56	[0.36, 0.76]	1.75
TAASSC	cl av deps	0.36	[0.12, 0.62]	1.44
TAASSC	acad lemma construction attested	0.21	[0.01, 0.42]	1.24
TAASSC	nsubj stdev	0.18	[0.01, 0.34]	1.20
TAASSC	av dobj deps	−0.21	[−0.40, −0.02]	0.81

Table 4: Cumulative models, focal predictors whose 95% CrI excluded zero, and the text length control (N = 700, eight-level outcome). All predictors were *z* standardized.

4.3 Boundary-Specific Effects

Table 5 reports the category-specific slopes from the three-band adjacent-category models (n = 618) at the lower boundary (A2|B1) and the upper boundary (B1|B2), for every predictor credible at one or both boundaries. Text length was credible at both boundaries for all four tools and was consistently stronger at the lower boundary than at the upper one.

Within TAALES, bigram association strength was credible at both boundaries (b = 1.32 and 1.10), whereas the age of acquisition of content words was credible only at the upper boundary (b = 0.72) and function-word concreteness only at the lower boundary (b = −0.68). Directional bigram association strength (COCA spoken bi DP), which was not credible in the cumulative model, was negative at the upper boundary (b = −0.58).

Within TAACO, subordinating conjunctions were credible at both boundaries (b = 0.50 and 0.35). Contrastive and adversative connectives were credible only at the lower boundary (b = 0.33), whereas semantic similarity to the source text (b = 0.64) and pronominal demonstratives (b = 0.44) were credible only at the upper boundary. The rate of source-text keyword use, which was not credible in the cumulative model, was negative at the upper boundary (b = −0.45).

Within TAALED, function-word diversity was credible at both boundaries with slopes of similar magnitude (b = 1.13 and 1.17), whereas content-word diversity was credible only at the upper boundary, where its slope was the largest observed in the analysis (b = 2.69).

Within TAASSC, the two boundaries were discriminated by different indices. Dependents per clause (b = 0.72) and the variability of dependents in subject noun phrases (b = 0.46) were credible only at the lower boundary, whereas complex nominals per clause (b = 0.70), the variability of dependents in object noun phrases (b = 0.57), words per clause (b = 0.37), and T-units per sentence (b = 0.33) were credible only at the upper boundary, as was the negative effect of dependents per direct object (b = −0.72). Words per clause, T-units per sentence, and the variability of dependents in object noun phrases had not been credible in the cumulative model.

Tool	Predictor	A2|B1	B1|B2
TAALES	COCA spoken bi MI	**1.32** [0.67, 2.00]	**1.10** [0.57, 1.67]
TAALES	Kuperman AoA CW	0.28 [−0.28, 0.84]	**0.72** [0.29, 1.16]
TAALES	Brysbaer _Concreteness Combined FW	**−0.68** [−1.16, −0.20]	−0.37 [−0.84, 0.09]
TAALES	COCA spoken bi DP	0.14 [−0.50, 0.79]	**−0.58** [−1.10, −0.05]
TAACO	lexical subordinators	**0.50** [0.13, 0.89]	**0.35** [0.003, 0.72]
TAACO	opposition	**0.33** [0.02, 0.65]	0.09 [−0.18, 0.36]
TAACO	source similarity word2vec	0.34 [−0.01, 0.73]	**0.64** [0.28, 1.00]
TAACO	unattended demonstratives	−0.05 [−0.42, 0.33]	**0.44** [0.10, 0.77]
TAACO	mag news uni keywords percentage	0.20 [−0.14, 0.55]	**−0.45** [−0.82, −0.09]
TAALED	hdd42 fw	**1.13** [0.46, 1.83]	**1.17** [0.56, 1.81]
TAALED	hdd42 cw	1.06 [−0.09, 2.25]	**2.69** [1.51, 3.85]
TAASSC	cl av deps	**0.72** [0.13, 1.32]	0.22 [−0.29, 0.73]
TAASSC	nsubj stdev	**0.46** [0.02, 0.89]	−0.24 [−0.57, 0.09]
TAASSC	CN/C	0.23 [−0.28, 0.73]	**0.70** [0.31, 1.09]
TAASSC	dobj stdev	−0.11 [−0.61, 0.38]	**0.57** [0.14, 1.00]
TAASSC	MLC	0.13 [−0.55, 0.91]	**0.37** [0.01, 0.80]
TAASSC	T/S	0.28 [−0.14, 0.72]	**0.33** [0.03, 0.65]
TAASSC	av dobj deps	−0.12 [−0.63, 0.37]	**−0.72** [−1.16, −0.31]

Table 5: Category-specific slopes (posterior mean [95% CrI]) at the two boundaries of the three-band adjacent-category models (n = 618). Only predictors credible at one or both boundaries are listed. Bold indicates that the 95% CrI excludes zero.

4.4 Model Comparison

In all eight comparisons of a null model with the corresponding main model, the main model predicted better, with elpd_diff ranging from −21.4 (SE = 8.1; TAASSC, cumulative) to −81.9 (SE = 12.2; TAALED, adjacent-category). Every difference was at least twice its standard error, so the improvement attributable to the focal predictors can be regarded as substantial throughout.

Within each family, the four main models were then

compared directly (Table 6). In the cumulative family TAALES ranked first, but the differences from TAALED (elpd_diff = −10.5, SE = 14.5) and TAACO (−13.1, 13.9) were within approximately one standard error, so these three tools could not be distinguished; TAASSC (−43.5, 12.9) fell short by approximately 3.4 standard errors and was clearly the weakest. In the adjacent-category family TAALED ranked first, and the differences from TAACO (−27.2, 14.3), TAALES (−36.2, 16.3), and TAASSC (−38.5, 18.3) were all approximately twice the corresponding standard error, indicating a clear advantage for TAALED. TAASSC ranked last in both families.

Family	Tool	elpd_diff	SE
Cumulative	TAALES	0.0	0.0
Cumulative	TAALED	−10.5	14.5
Cumulative	TAACO	−13.1	13.9
Cumulative	TAASSC	−43.5	12.9
Adjacent-category	TAALED	0.0	0.0
Adjacent-category	TAACO	−27.2	14.3
Adjacent-category	TAALES	−36.2	16.3
Adjacent-category	TAASSC	−38.5	18.3

Table 6: PSIS-LOO comparison across tools within each family, relative to the best-performing tool.

These cross-tool comparisons should be interpreted with some caution, since the four models were not matched in every respect: the number of focal predictors differed considerably across tools (from two for TAALED to sixteen for TAALES), the slope prior for TAACO was tightened relative to the other tools, and the screening thresholds were relaxed for TAALED and TAASSC. The resulting ordering across tools is therefore best read as suggestive of relative importance rather than as a strict ranking.

Finally, the two families were compared within each tool (Table 7). The adjacent-category model predicted better for three of the four tools, clearly so for TAALED (elpd_diff = +40.4, SE = 9.5) and TAASSC (+34.9, 16.5) but not clearly for TAACO (+15.8, 12.0); for TAALES the cumulative model was slightly better, again without a clear difference (−6.3, 13.6). The two tools with a clear advantage for the adjacent-category model were also those whose boundary-specific estimates in Table 5 diverged most clearly from the corresponding cumulative estimates in Table 4: words per clause and T-units per sentence in TAASSC, which were credible only at the upper boundary, and content-word diversity in TAALED, whose slope at the upper boundary was more than twice that at the lower boundary.

Tool	elpd_diff (adjacent − cumulative)	SE	Better family
TAALED	+40.4	9.5	Adjacent-category
TAASSC	+34.9	16.5	Adjacent-category
TAACO	+15.8	12.0	Adjacent-category
TAALES	−6.3	13.6	Cumulative

Table 7: PSIS-LOO comparison across families within each tool. Positive values favor the adjacent-category model.

5. Discussion

5.1 The Relative Importance of the Four Dimensions

In all eight comparisons between a null model and the corresponding main model, the selected indices improved predictive accuracy. Lexical sophistication, lexical diversity, cohesion, and syntactic complexity and sophistication therefore all carry information about CEFR level beyond the amount of speech. All the effects discussed below were estimated with text length held constant.

The four dimensions were not equally strong, however. In the cumulative family, TAALES, TAALED, and TAACO could not be told apart, and TAASSC was clearly the weakest. In the adjacent-category family, TAALED came first and TAASSC again came last. The strength of the lexical dimensions agrees with Lu (2012) and Bulté and Roothooft (2020), who also found lexical diversity to be important in rated speech. The weakness of the syntactic indices agrees with Iwashita et al. (2008) and Kolahi Ahari et al. (2025), who also found that syntactic measures separate levels of speaking proficiency poorly.

The adjacent-category model also predicted better than the cumulative model for three of the four tools, and clearly so for TAALED and TAASSC. The proportional odds assumption therefore does not hold for many of these indices. The sections below accordingly describe each dimension in terms of where on the CEFR scale its indices separate levels.

5.2 Lexical Sophistication and Diversity

Within lexical sophistication, the strongest predictor overall was bigram association strength measured by MI. It was also the only index in that set that was credible at both boundaries. CEFR level therefore depends not only on how difficult the individual words are, but also on how strongly the combined words are associated with each other. This agrees with earlier work showing that n-gram association strength is an important predictor of lexical proficiency beyond word frequency (Kyle et al., 2018). However, it should be noted that MI gives high values to low-frequency word pairs, and MI-based indices formed a separate dimension in Eguchi and Kyle (2020). The index may therefore partly reflect how rare the combined words are rather than how strongly they are associated.

The other lexical results divide the CEFR scale by word class: content words mattered at the upper boundary and function words at the lower boundary. The age of acquisition of content words was credible only at the upper

boundary, and the direction of its effect is worth noting: a higher age of acquisition went with a higher CEFR level. This is the opposite of the direction that Eguchi and Kyle (2020) reported for frequency-based sophistication in oral proficiency interviews. It is, however, the same direction that Kim et al. (2018) found in argumentative writing, where the components for content-word frequency and content-word properties were both negative predictors of proficiency. The direction of frequency effects and age-of-acquisition effects may therefore not depend on mode alone. The concreteness of function words was credible only at the lower boundary, and its effect was negative, which means that more abstract function words went with higher CEFR levels.

Lexical diversity showed the same division between the two word classes. Function-word diversity had almost the same slope at both boundaries ($b = 1.13$ and 1.17). Content-word diversity, in contrast, was credible only at the upper boundary, where its slope of 2.69 was the largest slope estimated in this study. This contrast also explains why function-word diversity had the larger coefficient in the cumulative model. Because that model imposes the proportional odds assumption and so captures the overall magnitude of an effect in a single coefficient, the consistency of function-word diversity's effect across both boundaries is what produced its larger cumulative coefficient. Content-word diversity's cumulative coefficient was smaller only because its effect was concentrated at the upper boundary; considered on this specific boundary, that boundary-specific effect was in fact considerably larger than that of function-word diversity, which illustrates the value of examining boundary-specific effects rather than relying on the cumulative coefficient alone.

This tendency was also visible in the learner responses. Lower-level speakers relied heavily on a relatively small set of highly frequent function words in simple patterns (e.g., *in*, *at*, *with*), typically to connect short clauses or indicate concrete locations (e.g., *in the library*, *at school*, *with my classmates*, *from northern Taiwan*). In contrast, higher-level speakers employed a wider range of function-word patterns, including more abstract relational expressions such as *instead of*, *regardless of*, *in terms of*, *rather than*, and *based on*. These expressions enabled speakers to express comparison, concession, cause, and abstract relationships beyond simple description.

The two lexical dimensions therefore point to the same developmental picture. On the lower boundary, learners are separated by their control of function words, both in how varied these words are and in how sophisticated they are. On the upper boundary, content words become more important, both in how varied they are and in how sophisticated they are. Previous studies have argued that content-word and function-word indices should be kept apart in studies of lexical sophistication in speech (Eguchi & Kyle, 2020; Eguchi, 2022). The present results show that the same argument applies to lexical diversity: combining the two diversity indices into a single score would have hidden the clearest level-dependent contrast in these data.

5.3 Source Text Integration

Among the cohesion indices, semantic similarity to the source text had the largest overall effect. In the three-band model it was credible only at the upper boundary. The fact that a measure of source-text integration predicts ratings agrees with the validation study for TAACO 2.0 (Crossley et al., 2019). It also agrees with earlier evidence that the degree to which speakers integrate source material predicts rated speaking proficiency (Crossley et al., 2014).

The second source-text index is more informative, because it behaved in the opposite way. The rate at which speakers repeated source-text keywords was not credible in the cumulative model, but it was negative at the upper boundary, exactly where semantic similarity was positive. The two indices measure similarity to the source in different ways. One compares the meaning of the whole response with the meaning of the whole source text. The other counts how many of the keywords of the source text appear again word for word. This difference suggests that paraphrasing the source-text keywords to some degree, rather than repeating them verbatim, is associated with higher ratings.

In Part 1, since the source text is not presented in the written form, the source text similarity might be influenced by learners' listening ability. On the other hand, in Part 2 and Part 3, the graphic information such as charts and figures as well as related questions and passages are displayed in written form, it is easier for speakers to refer to the source text while answering the questions. Moreover, in terms of proficiency levels, A2 level speakers tend to answer the questions with short answers while B1 and B2 level speakers use the rephrasing to gain thinking time and response immediately. Especially for Part 1, in which learners have to answer each question right after the question prompt, there is no time for learners to think and prepare their answers. Proficient learners usually use the rephrasing skills as one of the test-taking strategies, which leads to higher text similarity. By repeating the keywords and reconstructing the sentence in question prompts, speakers can start immediately and gain thinking time to finish the rest of the answer sentence. This also relates to the speaker's listening ability. Skilled listeners can identify the keywords and sentence structures while listening and verbalize the source text.

For example, in response to the question "*On average, how many hours do you spend studying every day?*", A2 level speakers answered only the number of the hours such as two hours or five hours, while B2 level speakers answered longer, using the keywords from the question prompt.

> *On average, I spend three hours every day on study because after uh learning at school I will always practice at home, so uh on average.*
>
> (S-2402_Part 1_Q1, B2)

5.4 Syntactic Complexity

The syntactic indices were the weakest set as a whole, but because TAASSC's adjacent-category model clearly outperformed its cumulative model in predictive accuracy,

the level-specific effects are especially important for interpreting TAASSC's results. Dependents per clause and the variability of dependents in subject noun phrases were credible only at the lower boundary. Complex nominals per clause and the variability of dependents in object noun phrases were credible only at the upper boundary. Elaboration at the clause level therefore separates the A2 band from the B1 band, and elaboration at the noun phrase level separates the B1 band from the B2 band. This pattern suggests that learners first increase syntactic complexity by expanding clauses rather than noun phrases. At the A2–B1 transition, learners appear to become able to elaborate their responses through subordinate clauses and other clause-linking devices, allowing them to express reasons, explanations, and additional information within a single utterance. Because these clause-level structures continue to be used at higher proficiency levels, however, they no longer distinguish B1 from B2. Instead, development at the upper boundary is characterized by increasing elaboration within noun phrases. Rather than simply producing objects such as *books* or *students*, B2 speakers more frequently produced expanded noun phrases.

> *So if stu-, if school only leave the places for central Taiwan student or for southern Taiwan student, it is very, it is uh, it is not an equal chance for northern Taiwan student to, to get for their academic, to get for their academic achievement.*
>
> (S-2404, Part 3, B2+ level)

> *And second question, I, although it although the passage is not totally correct but I still agree with the point that department make more electronic textbooks available to students and provide devices for reading.*
>
> (S2402, Part 3, B2 level)

The transition from B1 to B2 therefore appears to reflect not greater clause elaboration, but greater informational density within noun phrases.

This result relates to existing findings on two types of complexity: clausal complexity and phrasal complexity. Clause-level complexity has sometimes been characterized as a feature of spoken language, and phrase-level complexity as a feature of written academic language (Biber et al., 2011). The present data are entirely spoken, but phrasal complexity still developed in them, and it did so at the upper boundary. The contrast between clausal and phrasal complexity can therefore not be explained by mode alone (i.e. written vs spoken), because the two types appear at different points on the proficiency scale within a single mode.

One result seems to contradict this account. Dependents per direct object had a negative effect at the upper boundary, while complex nominals per clause had a positive effect at the same boundary. The two results are less inconsistent than they appear. Dependents per direct object is an average over all direct objects, including pronouns such as *it* and *him*, which cannot take modifiers. A speaker who uses many such pronouns will therefore have a low average, no matter how elaborate the other objects are. At the same boundary, the variability of dependents in object noun phrases was positive, which is what would be expected if elaborate objects and minimal objects were used in one response. Taken together, the two results fit speakers who treat their object noun phrases differently according to the context, choosing a pronoun when the referent is easy to recover and an elaborate phrase when it is not. They do not fit speakers who elaborate every object. The variability of dependents in subject noun phrases at the lower boundary can be read in the same way. It also suggests an order: variability appears first in subject position and later in object position. Speakers may become able to alternate between personal pronouns and complex nominals in subject position relatively early, but only become able to make the same distinction in object position at a later stage.

5.5 Limitations

There are three important limitations. The most significant one is that the eight independent responses were concatenated and analyzed as a single response. This matters in particular when interpreting the results of source-integration indices such as source text similarity. In the present study, all responses were combined into a single text so that all four tools could be compared on an equal footing, but future work will need to focus the analysis on each part separately in order to examine source integration in finer detail. The other limitation is that the boundary-specific effects are based on data that were collapsed and restricted to three proficiency bands. Because levels such as A1 and C1 were excluded, it was not possible to analyze which indices contribute to discrimination at the very lowest and highest levels. Future research should increase the sample size at the lowest and highest levels in order to analyze level-dependent patterns with greater precision.

A third limitation concerns the selection of predictors. Because the indices entered into each model were chosen on the basis of their rank correlation with CEFR, computed on the full dataset, the reported effect sizes and the improvement of each main model over its null model may be optimistically biased. In addition, the screening thresholds were relaxed for TAALED and TAASSC and supplemented by a manual review for TAACO, which increases the analytic flexibility of the procedure and limits its independent reproducibility. These issues bear more on the magnitude of the effects and on the ranking of the tools than on the direction of individual effects; the study's central finding — the point on the CEFR scale at which each dimension separates levels (Sections 5.2–5.4) — does not rest on the size of any single cumulative coefficient and is correspondingly more robust.

6. Conclusion

This study modelled CEFR speaking proficiency in a 700-response BESTEP corpus by relating indices of lexical sophistication, lexical diversity, cohesion, and syntactic complexity and sophistication to CEFR level through Bayesian ordinal regression. Once text length was controlled, all four dimensions carried information about proficiency, but they were not equally informative: the lexical dimensions were consistently among the strongest predictors, whereas syntactic complexity was the weakest. The adjacent-category models further showed that many indices do not separate the levels uniformly, and the boundary-specific analyses revealed a coherent

developmental picture in which function-word control and clause-level elaboration distinguish the lower levels, while content-word sophistication and diversity, source-text integration, and noun-phrase elaboration come to distinguish the higher levels. These level-dependent patterns are exactly the kind of criterial features that a CEFR-referenced test needs in order to justify its level boundaries, and they offer a first empirical basis for the eventual automated scoring of the BESTEP speaking test. These interpretations should nonetheless be read in light of the limitations set out in Section 5.5, whose resolution would allow the criterial features identified here to be specified with greater precision and validated for practical use.

7. Acknowledgements

This study was supported by an LTTC Teaching and Research Grant awarded to the "BESTEP Learner Corpus Development Project. (PI: Yukio Tono)"

8. Bibliographical References

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Bulté, B., & Roothooft, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, *91*, Article 102246. https://doi.org/10.1016/j.system.2020.102246

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Council of Europe.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages (CEFR): A manual*. Council of Europe.

Crossley, S. A., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*(3), 250–270. https://doi.org/10.1080/15434303.2014.926905

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*, 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192. https://doi.org/10.64152/10125/44329

Eguchi, M. (2022). Modeling lexical and phraseological sophistication in oral proficiency interviews: A conceptual replication. *Vocabulary Learning and Instruction*, *11*(2), 1–16. https://doi.org/10.7820/vli.v11.2.Eguchi

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381–400. https://doi.org/10.1111/modl.12637

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kolahi Ahari, M., Ghonsooly, B., Ghapanchi, Z., & Soodmand Afshar, H. (2025). Examining the role of lexical sophistication, lexical diversity, syntactic sophistication, syntactic complexity, and cohesion in L2 speaking proficiency assessment. *International Journal of Language Testing*, *15*(2), 43–70. https://doi.org/10.22034/ijlt.2025.492133.1395

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. https://doi.org/10.57709/8501051

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

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x

National Development Council, & Ministry of Education. (2021). *Bilingual 2030*. Executive Yuan, Taiwan.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4