Intelligibility and comprehensibility across Aptis Speaking tasks

Daniel R. Isbell[1], Dustin Crowther[1], Jieun Kim[1,2] & Yoonseo Kim[1]

[1]University of Hawai'i at Mānoa, USA

[2]Soongsil University, Republic of Korea

Current Aptis Speaking rubrics reflect Common European Framework of Reference (CEFR) Phonological Control descriptors, which were revised in 2018 to better emphasize the intelligibility of second language (L2) speech. The present study investigates the validity of Aptis Speaking scores and related CEFR-based interpretations by comparing official Aptis Speaking scores to laypersons' assessments of intelligibility, a measure of accuracy of understanding using orthographic transcription (0-100%), and comprehensibility, a measure of ease of understanding using ratings on a scale of 1 to 9. Additionally, as Aptis Speaking tasks feature several target performance levels, we considered relationships between task complexity and speakers' intelligibility and comprehensibility. Archived speaking performances from 50 Aptis examinees were assessed by layperson listeners for intelligibility (n = 562) and comprehensibility (n = 567). Comprehensibility was generally a stronger predictor of Aptis Speaking scores than intelligibility for both overall and task-level scores. Segmented regressions revealed specific breakpoints in which predictive power was greatest for each dimension (intelligibility up to 70% in transcription accuracy; comprehensibility ≥ 2). In sum, results from the current study generally provided support for the current Aptis Speaking rubrics, though the way in which comprehensibility is described at the upper levels may benefit from more nuance.

Keywords: Aptis, CEFR, comprehensibility, intelligibility, layperson listener

Email address for correspondence: disbell@hawaii.edu

© The Author(s) 2025. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits the user to copy, distribute, and transmit the work provided that the original authors and source are credited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

Pronunciation is a critical aspect of communicative speaking ability, as the phonological qualities of speech play a significant role in establishing listener understanding. As such, second language (L2) pronunciation scholarship has emphasized intelligible, easily comprehensible pronunciation as the preferred target of L2 pronunciation acquisition and teaching (Derwing & Munro, 2015; Levis, 2020). This preference is now better reflected in Common European Framework of Reference (CEFR) speaking descriptors, where an increased emphasis on intelligibility and comprehensibility over accentedness informed revisions to descriptors for Phonological Control (with a particular focus on the amount of effort required by an interlocutor to decode a message; Council of Europe, 2018). Similarly, scoring criteria for Aptis Speaking, which draw from CEFR level descriptors (O'Sullivan et al., 2020; see also Fairbairn & Dunlea, 2017), also feature pronunciation-related descriptors of speaking performance. Specifically, Aptis Speaking descriptors prioritize clarity of meaning and listener "strain" as opposed to nativelike production (see Aptis Speaking descriptors below). The current study aims to provide support for the validity of Aptis Speaking scores and related CEFR-based interpretations by comparing official Aptis Speaking scores and laypersons' assessments of speech. To do so we elicit laypersons' assessments of intelligibility, a measure of accuracy of understanding using orthographic transcription, and comprehensibility, a measure of ease of understanding using scalar ratings (Derwing & Munro, 2015). Given the graded nature of Aptis Speaking tasks, in which subsequent tasks are expected to elicit increased CEFR-linked target performance levels, we additionally consider task complexity, a factor found to influence L2 users' speaking production (Robinson, 2011; Skehan, 2009).

Literature review

Global perceptions of second language speaking

When Levis (2005) articulated a need to emphasize intelligibility over nativelikeness in L2 pronunciation development and teaching, he made clear that both speaker and listener were essential elements in the production of intelligible speech. This dual emphasis has

been well-documented in recent research on global dimensions of L2 speech, specifically in regards to Munro and Derwing's (1995) conceptualization of *accentedness* (i.e., how distinguishable an L2 utterance is from that of the target community), *comprehensibility* (i.e., perceived ease/difficulty of understanding an L2 utterance), and *intelligibility* (i.e., accuracy of understanding an L2 utterance). In reference to *listeners*, the emphasis has most frequently been placed on linguistic laypersons (sometimes referred to as naïve or untrained listeners), who are non-professionals without the training and metalinguistic knowledge of people typically employed to rate performances in high-stakes assessments (Sato & McNamara, 2019). Derwing and Munro (2009) argued that the perceptions of linguistic laypersons should serve as the gold standard for evaluating speech, as "what listeners perceive is ultimately what matters most" (p. 478). Layperson perceptions are particularly appropriate when investigating language in general, non-specialized contexts.

Referred to as global dimensions of L2 speaking, both accentedness and comprehensibility are typically assessed through scalar judgments (Derwing & Munro, 2015) while intelligibility is most commonly assessed through transcription (see Kang et al., 2018). Research has yielded consistent evidence that L2 speakers can produce speech that is both intelligible and comprehensible, even in the presence of a foreign accent (Crowther et al., 2022; Munro & Derwing, 1995). When rating accentedness, listeners mostly attend to phonological features of the speech stream, specifically the accuracy of individual sounds, but consider a wider range of phonological, lexical, and syntactic features when rating comprehensibility (e.g., Crowther et al., 2018; Trofimovich & Isaacs, 2012). Intelligibility has been associated with segmental accuracy, stress and intonation patterns, and repetition and pausing (e.g., Bent et al., 2007; Field, 2005; Nagle & Huensch, 2020; Nagle et al., 2023). In line with a shift away from nativelikeness towards general intelligibility, it has become common for L2 pronunciation research to prioritize both intelligibility and comprehensibility over accentedness, frequently excluding the latter altogether (e.g., see Crowther et al., 2022). This is the approach employed in the current study.

Task effects on global speaking performance

Task characteristics impact L2 speech production, affecting observable phonological, lexical, and grammatical aspects (Robinson, 2011; Skehan, 2009). How such task-based variation in speech production relates to listeners' subsequent perceptions is less clear. In a series of studies by Crowther (2020) and Crowther et al. (2015, 2018), listener-based comprehensibility and accentedness scores were higher in less complex tasks (as characterized by Robinson's 2011 Cognition Hypothesis), such as the IELTS long turn and an interactive opinion-sharing task, compared to more complex tasks, like a picture narrative and a TOEFL integrated speaking task. However, Crowther and colleagues' analyses of task differences were based on structurally different task types (e.g., picture narrative vs. opinion-sharing). Finer-grained test task manipulations, such as those employed by Aptis General, target clearly defined levels of task features such as communicative function, prompt modality, and response concreteness. This makes comparing task effects across various test and non-test sources challenging, and the trends observed by Crowther and colleagues require additional corroboration.

Those studies all focused on accentedness and comprehensibility; to our knowledge, no similar consideration has been given to task effects on the intelligibility of L2 speech. This is a noteworthy omission, as intelligibility is now strongly advocated for as the goal of L2 pronunciation acquisition and teaching (Levis, 2005, 2020), which is highly relevant given the extent to which L2 pronunciation continues to gain traction as a key criterion in L2 speaking assessment (see Isaacs & Harding's 2017 research timeline). One such high-stakes assessment in which this is evident is Aptis, which employs rubrics that directly refer to intelligibility (and comprehensibility) in line with the revised CEFR Phonological Control descriptors (Council of Europe, 2018).

Aptis Speaking Test

Aptis Speaking is one section of the four-skill Aptis General (interchangeably 'Aptis' hereafter), developed by the British Council (O'Sullivan et al., 2020). Aptis General serves as "the standard English language assessment product offered within the Aptis test system" (O'Sullivan et al., 2020, p. 5) and is promoted for use in a variety of settings, such

as universities and government. Development of the Aptis test system drew primarily on Weir's (2005) socio-cognitive framework, in which task demands are thought to interact with speakers' internal processes to influence response quality. Along these lines, test takers who differ in their processing capacity (i.e., in terms of linguistic resources and speech processing capacities) are expected to produce spoken responses that differ in quality. At the same time, a test taker's responses are expected to differ in accordance with task demands: less demanding tasks require less cognitive resource for reasoning/conceptualization and less strain on internal linguistic processing, leading to responses that demonstrate greater linguistic competence, while more demanding tasks stretch cognitive resources and internal processing, perhaps past limits, resulting in responses with more frequent linguistic breakdown. Accordingly, high-proficiency test takers are expected to demonstrate higher degrees of linguistic competence across tasks with varying demands, while lower-proficiency test takers may demonstrate higher degrees of competence in the context of less demanding tasks but struggle when task demands are greater.

Aptis Speaking performances are rated by trained human raters using task-specific scales which draw on CEFR performance standards, yielding two scores: a numerical score and a CEFR level classification. Aptis Speaking consists of four speaking tasks, each designed to elicit increasing target performance levels linked to CEFR levels: Short Response (A1/A2), Picture Description (B1), Picture Comparison (B1), and Long Response (B2). The demands of each task are raised by an increase in the number of communicative functions targeted, including informational and interactional features of speech. For example, in Task 1: Short Response, the primary informational features include *to provide personal information, to explain information/opinions*, and *to describe*. However, for Task 2: Picture Description, informational features additionally include *to elaborate* and *to justify an opinion*. Interactional features for Tasks 1–3 include *conversation repair*, but increase for Task 4 to also include *agreeing* and *disagreeing*. Table 1 elaborates on task differences along the lines of format, modality, concreteness of response, and communicative function. These differences in task design are meant to elicit higher levels of performance as test takers progress from Task 1 to Task 4.

Table 1. Aptis speaking tasks (O'Sullivan et al., 2020)

	Target CEFR Performance Level	Task Format	Modality of Prompts	Concreteness of Response	Communicative Functions	
					Informational	Interactional
Task 1: Short Response	A1/A2	Q&A	Verbal only	Highly concrete	Personal Info, Explain Opinions, Describe	Conversation Repair
Task 2: Picture Description	B1	Q&A	Verbal + Images	Fairly concrete	Task 1 + Elaborate, Justify Opinion	Conversation Repair
Task 3: Picture Comparison	B1	Q&A	Verbal + Multiple images	Fairly concrete	Task 1, 2 + Compare, Summarize, Express Preference	Conversation Repair
Task 4: Long Response	B2	Long-turn (with 60s planning)	Verbal + Picture	Fairly abstract	Task 1, 2, 3 + Staging	Conversation Repair, Agreeing, Disagreeing

Aptis Speaking descriptors

Aptis Speaking tasks are rated using holistic rubrics (available in O'Sullivan et al., 2020, pp. 90-92). With Aptis Speaking tasks designed to elicit increasing levels of spoken performance, the scale proficiency range from Task 1 through Task 4 differs. For Task 1, performance can be scored only up to B1, making it impossible to differentiate among B1, B2, and C level performances on this task alone. For Tasks 2 and 3, performance ratings can reach B2 level, though the lower end of the scale no longer differentiates A1 from no response. For Task 4, performance ratings can now reach C1 and C2 levels, with A1, A2, and no response undifferentiated.

As such, three distinct scales are employed for assessing Aptis Speaking performance that draw from CEFR assessment criteria, including Phonological Control (Council of Europe, 2018). Though speaking criteria additionally include grammar, vocabulary, and fluency, we focus on references throughout the three scales made to the intelligibility ("intelligible", "misunderstanding") and comprehensibility ("strain on the listener") of the speech produced, as highlighted in Table 2. By A2 level, test takers should be able to produce speech that is intelligible. While they are intelligible at A2, their production still likely places a strain on the listener, meaning that the listener needs to work harder to understand them (i.e., their speech is low in comprehensibility). While B1 and above performance is broadly intelligible, some degree of strain is still expected at B1 but absent at the B2 level and beyond.

Table 2. References to intelligibility and comprehensibility across Aptis Speaking Scales for tasks 1–4 associated with CEFR levels (emphases added: italics and underline for intelligibility, boldface for comprehensibility)

	Task 1	**Tasks 2 & 3**	**Task 4**
C2			Likely to be above C1 level.
C1			Uses a range of complex grammar constructions accurately. Some minor errors occur but **do not impede understanding**. *Pronunciation is clearly intelligible*.
B2		Likely to be above B1 level.	Some complex grammar constructions used accurately. **Errors do not lead to misunderstanding.** Sufficient range of vocabulary to discuss the topics required by the task. **Inappropriate lexical choices do not lead to misunderstanding.** *Pronunciation is intelligible.* **Mispronunciations do not put a strain on the listener or lead to misunderstanding.**
B1	Likely to be above A2 level.	Pronunciation is *intelligible* but inappropriate mispronunciations put an **occasional strain on the listener**.	Pronunciation is *intelligible* but occasional mispronunciations put an **occasional strain on the listener**.
A2	Mispronunciations are noticeable and **frequently place a strain on the listener**. Frequent pausing, false starts and reformulations but *meaning is still clear*.	Mispronunciations are noticeable and **put a strain on the listener.**	Performance is not sufficient for B1, or no meaningful language, or the responses are completely off-topic (memorised or guessing).
A1	*Pronunciation is mostly unintelligible* except for isolated words. Frequent pausing, false starts and reformulations **impede understanding**.	Performance below A2, or no meaningful language or the responses are completely off-topic (e.g., memorised script, guessing).	
Pre-A1	No meaningful language or all responses are completely off-topic (e.g., memorised script, guessing).		

Yan et al. (2018) investigated the intelligibility and comprehensibility of 125 Aptis Speaking Task 3 and 4 performances sampled from archived test data. For both dimensions there was a generally positive linear increase from examinees at the A1 through B1 levels, but little further increase in intelligibility or comprehensibility was seen among examinees beyond the B1 level (i.e., both dimensions leveled off). One limitation of Yan et al.'s analysis is the use of a subjective, scalar measure of intelligibility, which has been critiqued as being too similar to the scalar measure used to rate comprehensibility (Kang et al., 2018) and for lacking any means for listeners to demonstrate actual understanding. More commonly, intelligibility is measured through transcriptions, where listeners' recognition of words can be analyzed for accuracy. To our knowledge, however, transcription-based measures of intelligibility have not been applied to complete, extended speech samples. Further investigation employing common listener-based measures will help to determine what degree of accuracy is necessary at different levels for speech to be intelligible, and what degree of "strain" is acceptable for speech to be generally comprehensible.

The current study

Aptis General aims to assess "general proficiency" of adults relevant in a "wide range of general language use situations" (O'Sullivan et al., 2020, p. 17), and the Speaking scales emphasize intelligible, comprehensible speech through verbal descriptors that make clear reference to listener understanding and strain. Accordingly, investigating the extent to which official test scores align with the perceptions of intelligibility and comprehensibility by linguistic laypersons, who represent typical listeners in general real-world situations, would offer valuable validity evidence for the Aptis Speaking rating scales (Weir, 2005). In addition, knowing how the relationship between test scores and intelligibility and comprehensibility might be moderated by test task complexity would enrich the theoretical meaningfulness of scores. The following research questions guided our study:

1. What is the relationship between overall Aptis Speaking scores, intelligibility, and comprehensibility?

2. How do the relationships between Aptis Speaking scores, intelligibility, and comprehensibility vary across tasks?

For RQ1, we expected large, positive relationships between Aptis Speaking scores and laypersons' assessments of comprehensibility and intelligibility. As the CEFR Phonological Control descriptors from which Aptis draws its speaking criteria stress intelligible speech, we expect to see this reflected in the relationship between test scores and laypersons' perceptions. We expected comprehensibility and intelligibility in tandem to further account for variation in Speaking scores but had no strong hypothesis as to which dimension might prove to be more strongly linked to Speaking scores. For RQ2, we generally expected substantial relationships between scores on each Aptis Task and intelligibility and comprehensibility. We also expected lower comprehensibility and intelligibility on tasks targeting higher levels of performance and that this effect would be more pronounced for test takers with lower proficiency. Conversely, we expected that more proficient test takers would be more able to maintain high performance levels as task demands increase. Beyond these general hypotheses relevant to RQ1 and RQ2, we also sought a deeper understanding of how comprehensibility and intelligibility, in tandem and across the range of observed values, relate to proficiency scores, task ratings, and Aptis rubric descriptors.

Methods

This study used operational Aptis test data (speaking performances and scores) provided by the British Council. Anonymized study data is maintained by the Aptis Assessment Research Group, to whom requests for access may be directed. Statistical analysis scripts and a methodological appendix are openly available at https://osf.io/5gqds.

Aptis test data

We obtained archived speaking performances and scores from a sample of 50 Aptis General test takers stratified by overall Aptis score (10 each at CEFR A1-C levels) and representative of common test taker countries of residence in recent years (Vietnam = 15, Spain = 10, Albania = 5, Azerbaijan = 5, Colombia = 5, Mexico = 5, Saudi Arabia = 5).

These performances reflected the version of the test available in 2023. For each test taker, we were provided with four speaking samples, one per Aptis Speaking task. In total, 198 speaking samples were included in analyses, as two A1-level speakers did not include audible speech in one of their responses (one speaker missing Task 1, one missing Task 4).

All speech samples, originally collected in operational testing conditions, were edited to address issues with background noise, personally identifying information (e.g., stating one's name before starting a task response), and extended silence at the end of examinee responses. Samples were then transcribed, with unintelligible tokens recorded as /XXXX/. For comprehensibility ratings, complete samples for each task were used with a mean duration of 106.67s (n = 198, SD = 30.25s, range = 21.58s to 135.24s). For intelligibility ratings, the research team segmented each of the 198 transcriptions into a series of sequential segments based primarily on AS-units. An AS-unit is defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000, p. 365). AS-units longer than 24 words were further divided at a logical point to minimize listener difficulty. Segment length ranged from 2–24 words, with a mean length of 9.31 words. The resulting procedure yielded 2,846 intelligibility segments (mean # of segments = 14.23 per speaker per task, range = 2–33). Additional details on speech sample preparation are available in the Appendix (at the OSF link above).

Listeners

A total of 570 listeners were recruited via the online research participant pool *Prolific* (https://www.prolific.co), including 340 US-based (female = 156, male = 184) and 230 UK-based (female = 102, male = 128) listeners. The eligibility criteria for the listeners required them to be aged between 18 and 64 (mean = 36.50, SD = 10.6, min = 18, max = 64), use English as their primary language, have no reported hearing or language impairments, and be capable of touch typing. At the time of data collection, *Prolific* indicated there were 18,937 participants in the U.S. and 15,056 in the UK who met our eligibility criteria. *Prolific* distributed our study to all eligible participants, who could choose to participate on a first-come, first-served basis.

Instruments

Listeners judged comprehensibility on a 9-point scale (1 = 'hard to understand', 9 = 'easy to understand'; see Figure 1). Participants could select their rating at any time while listening but could not advance until the sound file had finished. Prior to judging, participants rated 3 practice speech samples and received feedback after each judgment (specifically, a verbal description, such as "highly comprehensible", and range of suitable ratings, such as "7, 8, or 9"). For intelligibility, participants transcribed entire task responses as a series of audio segments. For each segment, participants could begin typing the moment the sound file began to play (see Figure 2) and could advance to the next segment once the audio finished. In total, participants transcribed the speech of 4 speakers (1 task from each). Prior to beginning the intelligibility section, participants completed a 60-second typing test, consisting of a 100-word passage to reproduce, and then practiced transcribing the speech of 2 example speakers (no feedback was given). During training, participants were informed they could ignore punctuation, capitalization, and fillers.

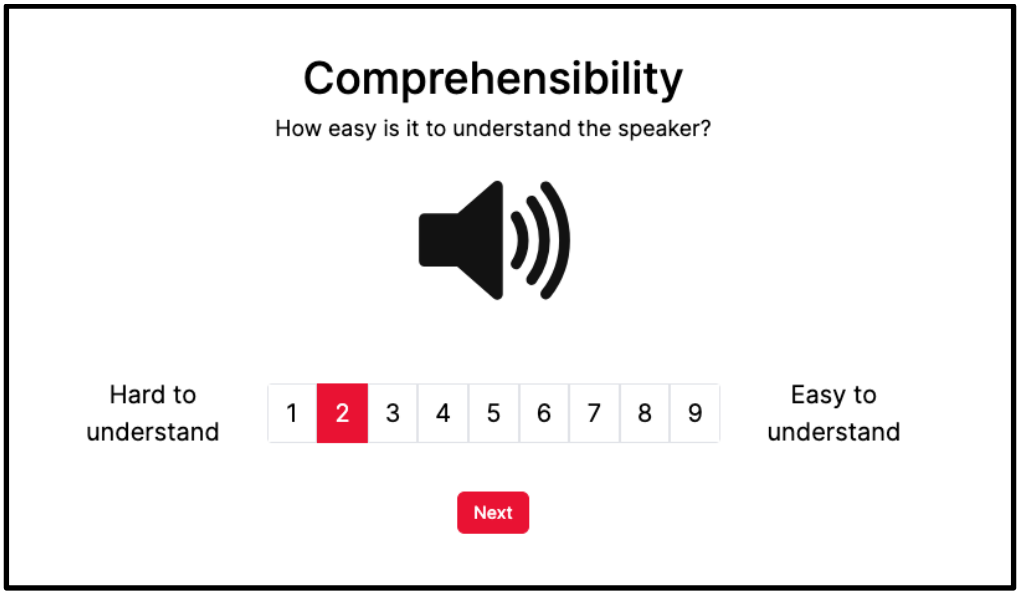

Figure 1. Comprehensibility rating screen, as presented on Gorilla

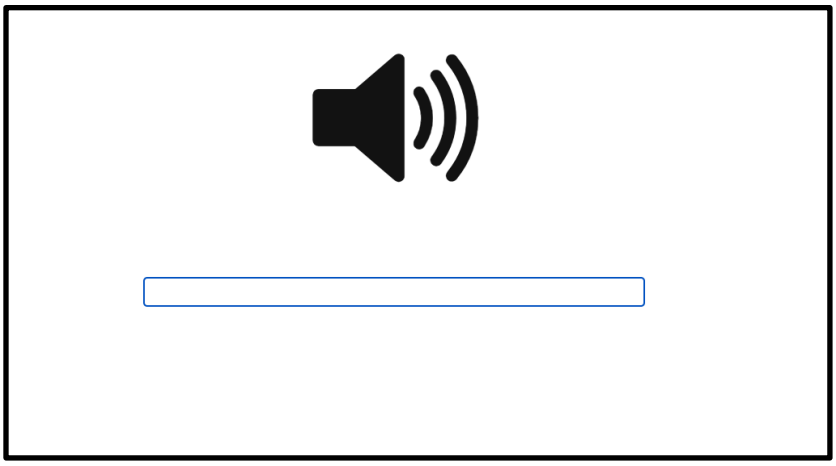

Figure 2. Intelligibility screen, as presented on Gorilla

Procedures

All data collection procedures were carried out via the online experiment platform Gorilla (Anwyl-Irvine et al., 2020). After they provided online consent and completed a background questionnaire, participants completed comprehensibility rating and intelligibility transcription procedures. Given the large number of speech samples to be rated and transcribed, a sparse rating design was employed (see Isbell, 2018). The 198 speech samples were divided into 4 blocks of 50 files each (A, B, C, D). To ensure enough overlap between listeners that would allow for the estimation of listener severity, the 4 blocks were then combined to form 4 supersets of speech samples (AB, BC, CD, DA). Each superset was then divided into 25 subsets of 4 files each, with each subset balanced for task, test taker's CEFR level, and test taker's L1. Each listener was then assigned to two supersets (e.g., AB & CD), and completed 1 subset within each (e.g., one subset within AB for comprehensibility ratings, one subset within CD for intelligibility transcriptions). Ultimately, each listener assessed one set of 4 speech samples for comprehensibility and transcribed another set of 4 speech samples for intelligibility, in that order. Aggregated at the speaker level (across all 4 tasks), there was an average of 45.3 judgments (*SD* = 2.02, min = 41, max = 49) per speaker on each dimension (intelligibility and comprehensibility).

Analyses

Data quality checks

Data quality is a universal concern in research, and especially in online data collection when researchers cannot directly monitor participants (Nagle, 2019; Nagle & Rehman, 2021). One participant's data was not successfully logged by Gorilla and could not be included. For comprehensibility ratings, we screened for uniform response patterns (e.g., a listener's comprehensibility ratings were identical for all speech samples) in the training task and the primary task that may have signaled low-effort or a misunderstanding of the scale. For intelligibility, we screened for listeners with large proportions of blank transcriptions (over 50% of assigned segments) and short average transcription length (< 20 characters) that might have indicated low effort or attention. Ultimately, we decided to exclude two listeners' comprehensibility ratings due to uniform response patterns, and seven listeners' transcription data, resulting in 567 listeners included for comprehensibility analyses and 562 for intelligibility. See the Appendix on the OSF site for additional details. For included listeners, typing test results indicated that the average typing speed was 48 wpm (*SD* = 19), with an average of 96% typing accuracy (cf. a mean of 40 wpm in Ostrach, 1997).

Scoring intelligibility responses

We scored transcriptions for accuracy by calculating the proportion (expressed as a percentage) of words in a listener's transcription which were in a criterion transcription. To create the criterion transcriptions, the research team first created verbatim transcripts through careful, repeated listening and reference to the Aptis speaking task prompts, which were provided by the British Council. Each verbatim transcript was confirmed by the second author. Tokens that were completely unidentifiable were recorded as /XXXX/. From these verbatim transcripts, a 'core' transcript which excluded extraneous elements, such as filled pauses (e.g., *um*), repetitions (e.g., *the the*), and self-corrections (with the correction being preserved). This was done to have the criterion transcription, and the evaluation of listener transcriptions, best reflect the speaker's message. For each segment, a listener's transcription was searched for a match with each word of the criterion

transcription. Then, the number of words correctly transcribed by each were aggregated across all segments of the task response to produce an accuracy score as a percentage. Transcription of excess words (e.g., transcriptions of repetitions or fillers) were ignored and did not impact transcription accuracy scores. Spelling mistakes were identified with the Google Sheets spellchecker, and corrected manually (e.g., “libary" changed to “library”).

Aggregating comprehensibility and intelligibility judgments

To arrive at aggregate measures of intelligibility and comprehensibility that accounted for differences in listeners’ ability to decode and transcribe speech (intelligibility) and severity of judgment (comprehensibility), we used many-facet Rasch measurement (MFRM) to analyze the listener response data in *Facets* (Linacre, 2021). We constructed three-facet models including speakers, listeners, and Aptis speaking task. For the comprehensibility model, a 9-point rating scale was specified. For intelligibility, we modeled the percentages using a scale based on 100 binomial trials. Logit-based estimates of speaker ability were transformed back to the original scale (1 to 9, or 0% to 100%) as *fair averages* that were adjusted to account for differences in other facets (i.e., listener severity).

To generate task-level fair averages, we estimated models using data only from a given task (i.e., two-facet models with speakers and listeners). Because there were not adequate links in the data when each task was analyzed separately, we anchored the listener measures from the aggregated models, allowing us to generate task-level speaker estimates based on known levels of listener ability/severity.

Relationships among intelligibility, comprehensibility, and Aptis scores

To address RQ1, correlational (Pearson and Spearman) and regressions analyses (linear and segmented, the latter with the *segmented* package, Fasola et al., 2018) were run in R. RQ2 required the use of Spearman correlations and ordinal regressions, given the nature of Aptis task-level scores. Proportional odds models with flexible (unstructured) thresholds for the response variable were run in using the *ordinal* package (Christensen, 2023). Additionally, for RQ2, descriptive statistics and data visualization techniques were

used to further understand the relationships between listener-based measures (intelligibility, comprehensibility) and Aptis Speaking task-level scores. Assumptions for correlation and regression models were checked visually (Lüdecke et al., 2021; Winter, 2020) and generally considered to be satisfied. One case appeared to be an outlier (a speaker with A2-level Speaking, but with very low comprehensibility and intelligibility) based on visual inspection and leverage statistics (i.e., Cook's distance) via the *performance* package (Lüdecke et al., 2021) in RQ1 and RQ2 analyses; however, removing this case did not substantially change results and it was therefore retained in reported analyses. For ordinal regressions, the proportional odds assumption was tested using the *scale test* function in the ordinal package; no violations were found.

Post-hoc exploratory analyses

Motivated by RQ1 and RQ2 findings, we conducted several post-hoc exploratory analyses to probe observed patterns in the data. This included data visualization and techniques to examine speaker consistency/variability across tasks, namely, the calculation of intraclass correlation coefficients (ICC), an index of consistency, and coefficient of variation (CV), which reflects variability as the speaker's standard deviation divided by their mean.

Results

Results related to Aptis Speaking scores (RQ1), task-level scores (RQ2) and post-hoc exploratory analyses are presented in the following subsections. First, Table 3 provides a descriptive summary of study variables. Comprehensibility values averaged around 4.50 in all tasks, and intelligibility values averaged approximately 72%.

Table 3. Summary of study variables

	n	*M*	*SD*	median	min	max
Overall						
Speaking Scale Score (0-50)	50	31.18	13.81	33.50	5.00	48.00
Comprehensibility (1-9)	50	4.54	2.15	3.90	1.19	8.60
Intelligibility (%)	50	72.06	15.13	75.65	20.06	89.85
Task 1						
Score (0-5)	50	3.76	1.48	4.00	0.00	5.00
Comprehensibility (1-9)	49	4.73	2.22	4.19	1.20	8.73
Intelligibility (%)	49	71.59	15.06	73.09	18.68	91.99
Task 2						
Score (0-5)	50	3.16	1.66	3.50	0.00	5.00
Comprehensibility (1-9)	50	4.54	2.22	3.96	1.15	8.78
Intelligibility (%)	50	73.06	14.87	75.78	20.83	94.16
Task 3						
Score (0-5)	50	3.22	1.56	3.50	0.00	5.00
Comprehensibility (1-9)	50	4.41	2.16	3.85	1.11	8.18
Intelligibility (%)	50	72.75	16.27	74.52	21.63	91.48
Task 4						
Score (0-5)	50	1.84	1.84	1.00	0.00	5.00
Comprehensibility (1-9)	49	4.63	2.07	4.56	1.31	8.59
Intelligibility (%)	49	72.47	13.99	73.32	34.06	92.89

Note. Comprehensibility and intelligibility values are computed as Rasch fair averages. Aptis scores were provided by the British Council.

RQ 1: Aptis Speaking scores, intelligibility, and comprehensibility

Pearson correlations between overall Aptis Speaking section scores and both listener-based measures were positive and large (Table 4), though the correlation with comprehensibility (r = .84) was noticeably larger than the correlation with intelligibility (r = .68). Based on Aptis Speaking section scores, test takers are assigned CEFR levels for speaking ability. The Spearman correlations (also in Table 4) between CEFR Speaking (an ordinal variable) and listener-based measures were also positive and large.

Table 4. Correlations between Listener-Based Measures and Aptis Speaking Scores

	Aptis Speaking					CEFR Speaking			
		95% CI					95% CI		
	r	LL	UL	*p*	R^2	ρ	LL	UL	*p*
Intelligibility	0.68	0.50	0.81	<.001	.46	0.72	0.54	0.83	<.001
Comprehensibility	0.84	0.73	0.90	<.001	.71	0.87	0.77	0.92	<.001

Note. *df* = 48 for all correlations.

Figure 3 shows relationships between Aptis measures of speaking proficiency (Aptis Speaking score, CEFR Speaking level) and the listener-based measures. For Aptis Speaking scores (Panels A and C), the loess curves (dashed lines) suggested that

relationships with listener measures are somewhat non-linear. The relationship between Aptis Speaking and intelligibility only began to increase sharply after about 60% intelligibility, while the relationship between Aptis Speaking and comprehensibility appeared stronger in the range of 1 to 5 than from 5 to 9. Panels B and D of Figure 3 are boxplots for listener measures at each CEFR Speaking level. The medians for intelligibility are further apart between lower CEFR levels (e.g., 13 percentage points between A1 and A2) and closer together at higher CEFR levels (e.g., 6 percentage points between B2 and C). It can be seen that most speakers at B1 or higher had aggregate intelligibility of 70% or more whereas A-level speakers had lower intelligibility (in some individual cases, much lower). Notably, a few A-level speakers had relatively high levels of intelligibility.

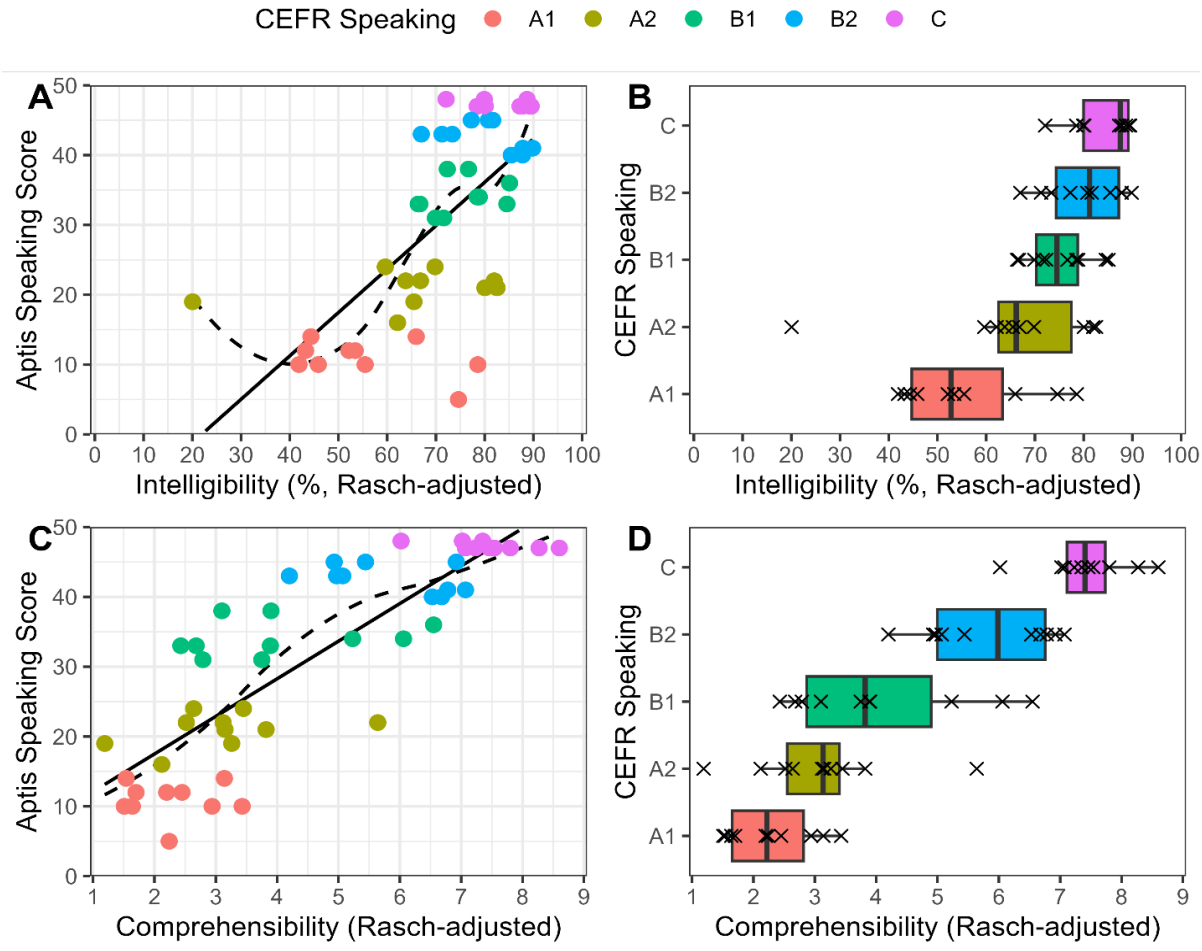

Figure 3. Relationships between Listener-Based Measures and (A, C) Aptis Speaking Scores and (B, D) CEFR Speaking Levels

To examine the associations with Aptis Speaking scores for intelligibility and comprehensibility in tandem, we ran two regression models (Table 5). The first model, a linear regression model with an R^2 of .70, suggested that comprehensibility had a much more substantive relationship with Aptis Speaking scores, with intelligibility accounting for no additional variation. The second model we ran was a segmented regression to account for potentially non-linear relationships between Aptis Speaking scores and the two listener-based measures. One breakpoint per predictor variable was deemed adequate based on visual inspection of the data and sample size considerations. In this model, the breakpoint for intelligibility was located at 70% and the breakpoint for comprehensibility was located at 1.72 on a scale of 1 to 9. For intelligibility, the slope before the breakpoint was positive, with a one percentage point increase in intelligibility associated with an increase of .59 Aptis Speaking points. After the breakpoint, the slope became -0.74. For comprehensibility, the slope before the breakpoint was extremely large and negative; the span of 1.00 to 1.72 only included 5 observations, however. After the breakpoint, the slope was large and positive, with a 1-point increase in comprehensibility associated with a 6.66-point gain in Aptis Speaking score. Notably, variance explained by the segmented model was 6-8% greater than that of the linear model. Substantively, it appeared that at very low levels of comprehensibility, intelligibility was a distinctive, positive predictor of Aptis Speaking scores, while comprehensibility itself was positively associated with Aptis Speaking scores throughout nearly the entire range of comprehensibility.

Table 5. Regressions of Aptis Speaking Scores onto Listener-Based Measures

	Linear Regression			Segmented Regression		
	Est.	95% CI	*p*	Est.	95% CI	*p*
Intercept	6.78	[-5.94, 19.50]	0.289	67.59	[7.19, 127.98]	0.029
Intelligibility	-0.00	[-0.25, 0.25]	0.992			
Breakpoint				70.02		
Slope 1				0.59	[0.07, 1.11]	0.028
Slope 2				-0.74	[-1.31, -0.18]	0.001
Comprehensibility	5.39	[3.61, 7.17]	<0.001			
Breakpoint				1.72		
Slope 1				-51.90	[-99.56, -4.24]	0.034
Slope 2				6.66	[4.72, 8.60]	0.017
R^2 / R^2 adjusted		0.701 / 0.688			0.779 / 0.748	

Note. *p*-values associated with Slope 2 terms are calculated in reference to a difference from Slope 1 terms.

RQ 2: Task effects on relationships among Aptis Speaking scores, intelligibility and comprehensibility

Aptis Speaking scores and listener-based speech measures were examined in further detail at the task level. Figure 4 illustrates the distribution of scores for each Aptis Speaking task, with corresponding CEFR levels noted. As seen in Table 3, higher numeric scores were seen in Task 1, and lower scores on Task 4, with Tasks 2 and 3 in between. For Task 1, the most common score was a 5, which corresponds to performance at the CEFR B1 level (or higher), while for Task 4 the most common score was a 0, reflecting performances at A2 (or below).

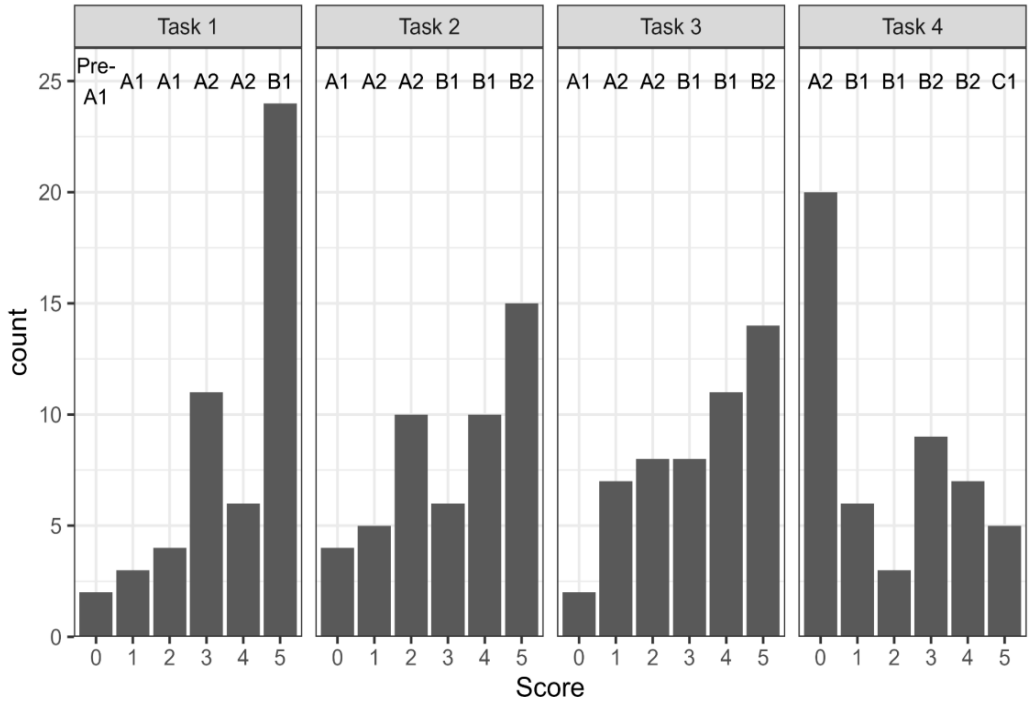

Figure 4. Histograms of Aptis Speaking Task Scores

It is important to note that task scores have ordinal properties, in line with the CEFR. Accordingly, we used Spearman correlations and ordinal regressions to model how intelligibility and comprehensibility predicted task scores, both separately and in tandem. As shown in Table 6, Spearman correlations were statistically significant for intelligibility and comprehensibility across all tasks. Intelligibility correlations were medium to large,

while comprehensibility correlations were consistently large and were stronger than intelligibility correlations for each task.

Table 6. Spearman Correlations between Aptis Task Scores and Listener Perceptions

		Intelligibility				Comprehensibility			
			95% CI				95% CI		
	n	ρ	LL	UL	*p*	ρ	LL	UL	*p*
Task 1	49	0.47	0.21	0.67	<.001	0.69	0.49	0.81	<.001
Task 2	50	0.70	0.52	0.82	<.001	0.80	0.67	0.88	<.001
Task 3	50	0.59	0.36	0.75	<.001	0.77	0.62	0.87	<.001
Task 4	49	0.57	0.34	0.74	<.001	0.75	0.58	0.85	<.001

To examine the relationships intelligibility and comprehensibility had with Aptis Speaking task scores in tandem, ordinal regression models were run for each task (Table 7). These models explained between 56% and 59% of variance in the task scores. For all tasks, only comprehensibility was a statistically significant predictor, with a 1-point increase in comprehensibility associated with substantially higher odds of earning a higher task score after accounting for intelligibility. In contrast, intelligibility appeared to have little to no influence on Aptis Speaking task scores after accounting for comprehensibility.

Table 7. Summary of ordinal regressions of Aptis Speaking Task Scores onto Listener-Based Measures

		Intelligibility + Comprehensibility		
	n	*OR*	*p*	R^2
Task 1	49			.59
Intelligibility		.94	.052	
Comprehensibility		3.66	<.001	
Task 2	50			.59
Intelligibility		1.00	.983	
Comprehensibility		2.81	<.001	
Task 3	50			.58
Intelligibility		.98	.498	
Comprehensibility		2.97	<.001	
Task 4	49			.56
Intelligibility		1.01	.844	
Comprehensibility		2.46	<.001	

Note. OR = Odds ratio, R^2 = Nagelkerke's R^2.

Post-hoc exploratory analyses

To further probe the relationships between tasks and speakers' performances, we carried out descriptive graph-based analyses and quantitatively examined speaker

consistency/variability across tasks in terms of intelligibility and comprehensibility. Figure 5 shows mean listener-based measures of speakers at different overall CEFR Speaking Levels for each task. Due to the small *n*-size in CEFR-based speaker subgroups (n = 10 per group), caution is needed when interpreting these results. For both intelligibility and comprehensibility, means follow an expected order for each task, with C-level speakers at the top and A1-level at the bottom. However, as the 95% confidence intervals illustrate, it is comprehensibility scores that more clearly differentiate speakers at adjacent proficiency levels. While C-level speakers clearly demonstrate higher levels of intelligibility than A-level speakers, their superiority to B2-level speakers is not statistically significant for all tasks, as shown by substantially overlapping confidence intervals (Tasks 2, 3, and 4). For both intelligibility and comprehensibility, certain tasks seemed to help distinguish speakers at adjacent CEFR levels. For example, the gaps between A1 and A2 speakers widened on Task 4 for both measures, and the difference between B1 and B2 also seemed to be magnified by Task 4.

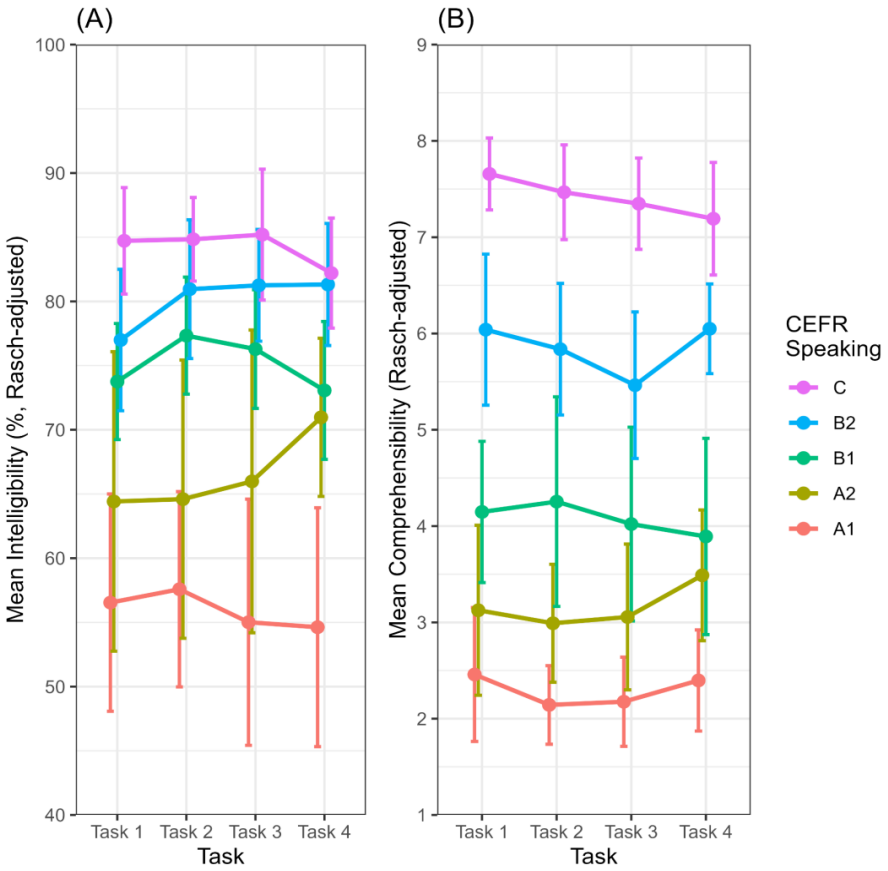

Figure 5. (A) Intelligibility and (B) Comprehensibility across Aptis Tasks by CEFR Speaking Level
Note. Error bars represent 95% confidence intervals.

Finally, we examined the relationship between task-level performance and listener-based measures among speakers at different overall CEFR Speaking levels. Figure 6 shows intelligibility and comprehensibility measures across tasks for individuals, grouped by overall CEFR level. Notably, there was considerable within-speaker stability across tasks, corroborated by intraclass correlation (ICC) values of .91 for intelligibility and .94 for comprehensibility. For intelligibility, speakers within the same CEFR Speaking level appeared tightly clustered (except A1) and adjacent levels demonstrated considerable overlap. Speakers at different CEFR levels were more clearly distinguished by comprehensibility, though some overlap at adjacent levels was visible, particularly at the A2 and B1 levels. Additionally, speakers with higher intelligibility and comprehensibility demonstrated less variation in intelligibility/comprehensibility across the four Aptis tasks, as indicated by medium-to-large negative correlations between mean listener-based measures and coefficients of variation (intelligibility: r = -.63 [-.77, -.42], p <.001; comprehensibility: r = -.55 [-.71, -.31], p < .001).

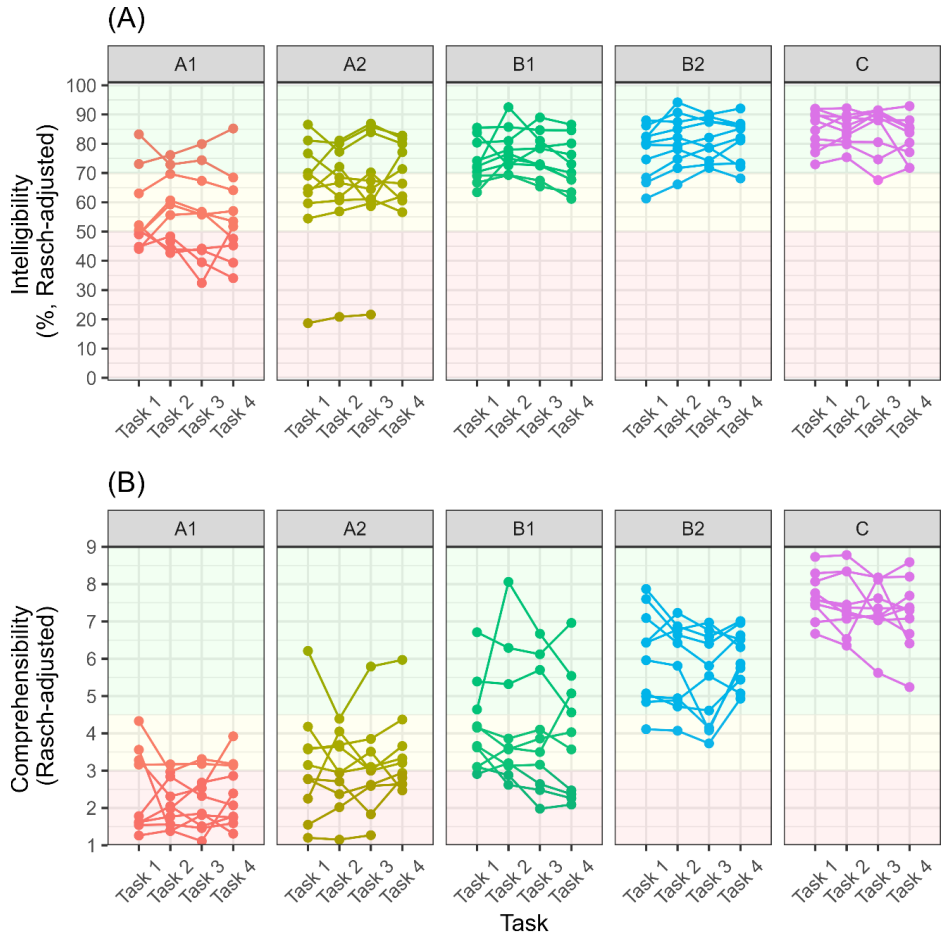

Figure 6. Individual (A) Intelligibility and (B) Comprehensibility Scores across Aptis Speaking Tasks

Discussion

In this study, we set out to examine the relationship between Aptis Speaking scores and two constructs seen as critical to the speaking construct, intelligibility and comprehensibility, that are embedded in Aptis rubrics. In the following discussion, we consider the answers to the study's research questions and implications for the Aptis test.

RQ 1: What is the relationship between overall Aptis Speaking scores, comprehensibility, intelligibility?

The relationships between Aptis Speaking score and comprehensibility (r = .84) and intelligibility (r = .68) were both strong. Moreover, a segmented regression model including intelligibility and comprehensibility as predictors explained 75% of variance in Aptis Speaking scores. For comprehensibility, past a breakpoint of 1.79 (on a 9-point scale) each 1-point increase in comprehensibility was associated with a substantial increase (6.66 points, out of 50 possible) in Aptis Speaking score. Interestingly, intelligibility was a positive predictor of Aptis Speaking score only up to 70% intelligibility, with a 10-percentage point increase associated with a 5.9-point increase in Aptis Speaking score. These general results provide some strong support for the meaningfulness of Aptis Speaking scores: to the ears of layperson listeners (Sato & McNamara, 2019), speakers that earned higher Aptis scores were more intelligible and easier to understand.

Given that comprehensibility alone correlated with Aptis Speaking scores very strongly, with an R^2 of .71, it would seem clear that comprehensibility was overall the most important predictor of Aptis Speaking scores – particularly for test takers who are able to maintain modestly high levels of intelligibility. Of course, as even the segmented regression model with both intelligibility and comprehensibility as predictors explained only about 75% of variance (conservatively, based on adjusted R^2) in Aptis Speaking scores, the remaining 25% of variance is attributable to other factors, with likely candidates being other elements featured in the Aptis rubric descriptors like grammatical complexity and accuracy, lexical selection, fluency of speech (all of which have been shown to also be predictors of comprehensibility; e.g., Crowther et al., 2018; Trofimovich & Isaacs, 2012), and topic fulfillment.

Test takers in the B1, B2, and C levels generally produced speech beyond (or very close to, in some cases) the 70% intelligibility breakpoint, and test takers at the A1 level were mostly below this breakpoint (A2 test takers were split below and above; see Figure 3). This pattern seems to align with Yan et al. (2018), who found that intelligibility differences from B1 to C levels were minimal among Aptis examinees. For comprehensibility, there were clearer distinctions between C-level and B2-level speakers, and between B2-level and B1-level speakers. Though still ordered in line with expectations (B1 > A2 > A1), more overlap in comprehensibility was seen between the three lower levels. These findings differ from Yan et al. (2018), where comprehensibility differences among higher CEFR-level examinees were minimal. Overall, the present findings seem reasonably in line with the updated CEFR descriptors for Phonological Control, which note limitations to intelligibility at lower levels (particularly A1 and A2). The present cross-sectional findings also corroborate some key claims made by L2 pronunciation researchers about intelligibility and comprehensibility (Derwing & Munro, 2015; Levis, 2020), including the partial independence of the two constructs and a general level of intelligibility as a requirement for speech to be comprehensible.

RQ 2: How do the relationships between Aptis Speaking scores, intelligibility, and comprehensibility vary across tasks?

Comprehensibility had stronger correlations with task scores than intelligibility for all tasks. Additionally, ordinal regression models with both intelligibility and comprehensibility included as predictors suggested that intelligibility contributed little to predictions of task scores after accounting for comprehensibility. Interestingly, bivariate correlations for both intelligibility and comprehensibility were strongest for Task 2, Picture Description. Noting that this difference was not especially large, we nonetheless speculate that it may be due to the specific lexical demands of the task: Knowing how to intelligibly (and comprehensibly) pronounce words associated with a given picture may be especially critical for successful task completion.

We originally hypothesized that more capable test takers would differentiate themselves from less capable test takers as task complexity increased. Strong ICCs at the speaker level across tasks suggested that relative performance, in both comprehensibility and

intelligibility, was consistent across tasks, which is evidence against the hypothesis. Some evidence for the predicted pattern, however, was visible at the level of CEFR-based speaking proficiency groups, particularly in comprehensibility, with some tasks obscuring or magnifying between-level differences. C-level test takers were consistently rated as more comprehensible than B2-level test takers across tasks, while B2 speakers differentiated themselves more from B1 on Tasks 1 and 4. B1 speakers substantially overlapped with A2 speakers, who in turn overlapped with A1 speakers. Task 4 appeared to magnify the differences between A1-A2 and B1-B2 test takers. For intelligibility, there was substantial overlap between neighboring levels, except for A1-A2 on Task 4.

Although within-speaker stability was strong, speakers' intelligibility and comprehensibility did fluctuate across tasks. There were test takers whose intelligibility and comprehensibility decreased steadily from Task 1 through to Task 4; however, there were also test takers who demonstrated a gain in one of these speech dimensions from one initial task to a later, more demanding task (Figure 6). This inconsistency requires further consideration, and may have to do with the use of a holistic rubric and the quality of task completion: A speaker who failed to effectively complete all aspects of a more demanding task may have been able to produce intelligible/comprehensible speech in whatever it was they did produce in their response, but ultimately earned a low mark.

Implications for Aptis Speaking descriptors

Aptis Speaking rubrics were developed with CEFR Phonological Control descriptors in mind (Council of Europe, 2018), frequently referencing both intelligibility and comprehensibility (i.e., "strain"). The perceptions of layperson listeners in this study, with no formal training on the Aptis Speaking rubrics, generally align with the descriptors. In addition, these laypersons' responses can also help refine the qualifiers of intelligibility and comprehensibility at different levels of performance.

Beginning with intelligibility, transcription accuracy up to 70% was found to predict Aptis Speaking scores. Only A1 test takers, at a level in which intelligibility is not assumed within the Aptis Speaking rubric, consistently performed below this point. While a few A1 speakers registered intelligibility below 50%, most produced speech that was more

intelligible than not (> 50%). It may be the case that "mostly unintelligible" in the Aptis rubric is too strict a criterion. All other levels either hovered around (A2) or mostly fell above (B1, B2, C) the 70% threshold. While there was indeed incremental growth in intelligibility scores from A2 through C, there was also considerable overlap between levels. Overall, intelligibility offering little to no prediction of Aptis Speaking scores past A1 reflects the current Aptis Speaking rubrics.

Comprehensibility serves as a stronger predictor of beyond-A1 performance. This also reflects the current Aptis Speaking rubrics, in which comprehensibility goes from "strain" (A2) to "occasional strain" (B1) to "do not put a strain on the listener" (B2 and above). By employing the ratings of laypersons, our data provides some nuance to these descriptions. Referring to the visual results in Figure 3, A1 test takers, who we have already designated as notably unintelligible, have comprehensibility ratings, on average, < 3. For A2, where frequent "strain" exists, comprehensibility ratings fell between 2–4. For B1, with "occasional strain", comprehensibility ratings were more variable, but generally fell between 3–5. Interestingly, while the current Aptis Speaking descriptors indicate that there is "no strain" at the B2 and C levels, our data allows us to still differentiate between the two levels. B2 test takers have comprehensibility generally between 5–7, while C test takers are ≥ 7. Following laypersons' ratings of comprehensibility, 'no strain' as indicated by current Aptis Speaking descriptors would be any rating above 5. However, given the evident difference in comprehensibility judgments between B2 and C, current descriptors may benefit from minor revision. One potential approach might be the following: B2 as "limited strain", C as "very little to no strain". As no test takers, regardless of their Aptis Speaking score, achieved an aggregate comprehensibility rating of 9, any descriptor amounting to 'no strain' may be unrealistic.

Limitations and future directions

While our stratified sample of 50 Aptis test takers allowed us to consider performance across a range of proficiency, the smaller subgroup sizes corresponding to CEFR levels limited several potential takeaways to descriptive and visual considerations. Specifically, a larger sample per CEFR level would allow for more rigorous investigation into visual patterns identified for RQ2, which investigated task effects on the relationship between

Aptis Speaking score and intelligibility and comprehensibility. In addition, we recruited listeners through *Prolific*, an online panel. While the use of such resources has increased in L2 research (see Nagle, 2019; Nagle & Rehman, 2021), limitations in terms of listening/rating environment standardization should be acknowledged.

Levels of intelligibility (M = 72%) observed in this study are lower than in some previous research on L2 speech (e.g., Huensch & Nagle, 2021, where most utterances were transcribed with near-perfect accuracy). Three factors might explain this outcome: (1) sampling speakers from a wide range of proficiency, (2) drawing on extended, spontaneous speech samples in their entirety, and (3) adverse listening conditions linked to audio file quality or listening environment. All three factors are worth exploring, although we wish to elaborate specifically on (1) and (2). First, we purposefully included test takers representing a full range of proficiency (A1 through C). Previous intelligibility research has frequently worked with a more constrained sample in terms of proficiency. For example, Munro and Derwing (1995) featured speakers at the equivalent of B2 or above, with ~80% of speakers transcribed at 80% accuracy or higher. It is possible that our lower mean intelligibility scores stem from lower proficiency participants in the sample. However, Huensch and Nagle (2021), who reported near-perfect accuracy in transcriptions, featured university students with a large range of proficiency as measured on an elicited imitation test. Second, to our knowledge, no previous study has attempted to analyze intelligibility over extended, spontaneous speech. In most intelligibility research, listeners are provided with decontextualized utterances drawn from a larger response (see Kang et al., 2018), which allows researchers to better control the language to be transcribed. In the current study, focused on the relationship between intelligibility and Aptis Speaking scores, we felt it necessary to ensure that listeners were exposed to the entirety of test task responses. This meant that utterances were full of repetitions, false starts, and filler, frequently featured ungrammatical structures, and ranged from relatively short (2 words) to relatively long (24 words) segments. As segment length has been shown to predict intelligibility accuracy (e.g., Nagle & Huensch, 2020), this may be another reason for the lower intelligibility scores. The use of chronological, contextualized segments as a measure of intelligibility needs further methodological investigation.

Conclusion

This study, through investigating relationships between laypersons' ratings and Aptis Speaking scores, provides useful elaboration on what it means to be intelligible and comprehensible, in quantitative terms, at different levels of Aptis Speaking performances. While providing support for current intelligibility- and comprehensibility-oriented descriptors across the three rubrics used to assess Aptis Speaking performances, this study's findings also suggest avenues for revisions that would more accurately characterize A1-level intelligibility and better characterize and differentiate comprehensibility performance at the B2 and C levels. More broadly, this study has yielded a deeper understanding of how listener-based measures of intelligibility and comprehensibility interface with speaking proficiency, supported by a novel approach of measuring intelligibility in extended discourse. Findings that it is comprehensibility of speech that distinguishes more proficient speakers past a moderately high threshold of intelligibility (roughly 70% in this study) offer support for mainstream theoretical frameworks (Derwing & Munro, 2015; Levis, 2020) and hold implications for the development of assessments (i.e., in terms of rubrics and automated scoring engines) and language curricula (in terms of content and sequencing of pronunciation-focused teaching) alike.

Author disclosures

In the last five years, D. Isbell and D. Crowther have received research funding from Duolingo and Educational Testing Service. D. Isbell has received research funding from Pearson. D. Isbell has received honoraria from Educational Testing Service; consulting fees from Cambridge University Press and Assessment, IELTS UK, and Duolingo. J. Kim has received research funding from British Council, Duolingo, and Educational Testing Service. The authors have no other potential conflicts of interest to report.

This research was funded by a British Council Assessment Research Grant awarded to Dustin Crowther (Principal Investigator) and Dan Isbell (Co-Investigator).

The respective roles of authors are as follows:

D. Isbell: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing

D. Crowther: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing

J. Kim: Formal analysis, Investigation, Writing – original draft, Writing – review & editing

Y. Kim: Formal analysis, Investigation, Writing – original draft, Writing – review & editing

ORCID iDs

Daniel R. Isbell https://orcid.org/0000-0001-5850-8809

Dustin Crowther https://orcid.org/0000-0002-4399-0169

Jieun Kim https://orcid.org/0000-0001-5924-370X

Yoonseo Kim https://orcid.org/0000-0003-0566-8715

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bent, T., Bradlow, A. R., & Smith, B. L. (2007). Segmental errors in different word positions and their effects on intelligibility of non-native speech. In O. S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning:*

In honor of James Emil Flege (pp. 331–347). John Benjamins. https://doi.org/10.1075/lllt.17.28ben

Christensen, R. (2023). *ordinal-Regression models for ordinal data*. R package version 2023.12-4, https://CRAN.R-project.org/package=ordinal

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. *Companion volume*. Council of Europe. www.coe.int/lang-cefr

Crowther, D. (2020). Rating L2 speaker comprehensibility on monologic vs. interactive tasks: What is the effect of speaking task type? *Journal of Second Language Pronunciation*, *6*(1), 96-121. https://doi.org/10.1075/jslp.19019.cro

Crowther, D., Holden, D., & Urada, K. (2022). Second language speech comprehensibility. *Language Teaching, 55*(4), 470-489. https://doi.org/10.1017/S0261444821000537

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal, 99*, 80-95. https://doi.org/10.1111/modl.12185

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*(2), 443-457. https://doi.org/10.1017/S027226311700016X

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476–490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins. https://doi.org/10.1075/lllt.42

Fairbairn, J., & Dunlea, J. (2017). *Technical report: Speaking and writing rating scales revisions*. British Council.

Fasola S., Muggeo, V.M.R., & Kuchenhoff, K. (2018). A heuristic, iterative algorithm for change-point detection in abrupt change models. *Computational Statistics, 33*, 997-1015. https://doi.org/10.1007/s00180-017-0740-4

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*(3), 399-423. https://doi.org/10.2307/3588487

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375. https://doi.org/10.1093/applin/21.3.354

Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning, 71*(3), 626–668. https://doi.org/10.1111/lang.12451

Isaacs, T., & Harding, L. (2017). Pronunciation assessment. *Language Teaching*, *50*(3), 347-366. https://doi.org/10.1017/S0261444817000118

Isbell, D. R. (2018). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 89–111). Routledge. https://doi.org/10.4324/9781315170756-6

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the Intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*, 115-146. https://doi.org/10.1111/lang.12270

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369-377. http://www.jstor.org/stable/3588485

Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, *6*(3), 310-328. https://doi.org/10.1075/jslp.20050.lev

Linacre, J. M. (2021). *A user's guide to FACETS Rasch-model computer programs* (program manual 3.83.5). Winsteps.com.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, *5*(2), 294-323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C. L., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*, *6*(3), 329-351. https://doi.org/10.1075/jslp.20009.nag

Nagle, C. L., Huensch, A., & Zárate–Sández, G. (2023). Exploring phonetic predictors of intelligibility, comprehensibility, and foreign accent in L2 Spanish speech. *The Modern Language Journal*, *107*(1), 202-221. https://doi.org/10.1111/modl.12827

Nagle, C. L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, *43*(4), 916-939. https://doi.org/10.1017/S0272263121000292

Ostrach, T. R. (1997). *Typing speed: How fast is average: 4,000 typing scores statistically analyzed and interpreted*. https://www.scribd.com/doc/62673503/Typing-Speed

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K. (2020). *Technical report: Aptis general technical manual* (version 2.2). British Council. https://www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf

Robinson, P. (Ed.). (2011). *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*. John Benjamins. https://doi.org/10.1075/tblt.2

Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics, 40*(6), 894–916. http://doi.org/10.1093/applin/amy032

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532. https://doi.org/10.1093/applin/amp047

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905-916. https://doi.org/10.1017/S1366728912000168

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Winter, B. (2020). *Statistics for linguists: An introduction using R*. Routledge. https://doi.org/10.4324/9781315165547

Yan, X., Kim, H. R., & Kim, J. Y. (2018). Complexity, accuracy and fluency (CAF) features of speaking performance on Aptis across different levels on the Common European Framework of Reference for Languages (CEFR). *ARAGs Research Reports Online*. https://www.britishcouncil.org/sites/default/files/yan_et_al_b.pdf