Envoy
Design Manual

Draft V 8.4.
Last updated 5 April 2024

This is a live working document intended to provide information about the design process of the Envoy English test. Information contained herein is subject to change. Updated versions of this document will be available on the Envoy website.

Contents

Abstract

Envoy is a new English language proficiency test. It measures foundational skills, language reception, and language production in academic and general (daily life) contexts. The test covers the four skills of reading, listening, speaking, and writing, and additionally provides information about the test taker's knowledge of grammar, vocabulary, and cohesion. This paper presents the concepts underpinnings and development process of Envoy. The purpose and intended uses of the test, its target test-taker population, and relevant language use domains are described first. This paper outlines the rationale, intended applications, and demographic focus of the Envoy test, delineating its relevance across various domains of language use. Subsequent sections detail the test's design and scoring procedures, culminating in a description of the systematic approach employed for ongoing research and validation of the exam.

General Description of Envoy

Envoy is a Common European Framework of Reference (CEFR) aligned computer-based test of English proficiency that includes sections on language knowledge, language reception skills, and language production skills. The test covers the four language skills of listening, reading, speaking, and writing, and measures knowledge about language through sections on grammar, vocabulary, and cohesion. The skills sections are modular and can be administered in any combination. The test requires approximately 90 minutes to complete if all modules are included. The test is adaptive and is suitable for language learners from CEFR level A1 through to C2.

The test integrates Artificial Intelligence (AI) in proctoring and scoring, which enables unbiased, consistent, and rapid results within 5 hours. Human raters are used to fine tune the AI scoring. Human raters also review tests flagged for misconduct by the AI.

Envoy can be taken on any desktop computer or laptop that is connected to the internet and has a functioning camera and microphone.

Envoy focuses on communicative skills through use of authentic, natural language. Listening and reading passages are drawn from authentic materials covering all difficulty levels from the B1 level and above. Passages are drawn from academic and daily life domains. Speech and writing are measured through open-ended questions.

The purpose of the test is to assess English language proficiency, either as a stand-alone English proficiency test, or for use in English-language courses and programs. Within these programs, the test is useful for placement, progress assessment, or as an exit requirement. The test is appropriate for all levels of language learners and integrates both academic and daily domains of language. To ensure the generalizability of results, the test is aligned to the CEFR.

Validity

Sections

Prior to the launch of Envoy, a pilot was conducted with 150 students from several different countries. Descriptive and analytical statistics were collected for all complete tests (n=96) to analyze internal consistency of the test sections. Statistics were collected on the mean score for each section, the standard deviation, the standard error of means, and the confidence intervals.

Analysis of the correlation between the different sections showed that the sections were all correlated with each other, but not to the extent that would indicate that sections are testing the same skills. The standard error of measurement was under 0.2 points for each section, and did not differ significantly among sections, indicating that the sample scores are predictive of the population scores.

Levels

Construct validity is "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1996, p. 231). In this case, the claim is that the test measures the level of competence of the test taker for the purposes of successful communication in a professional or academic context. Part of the validity of Envoy is obtained through alignment with the Council of Europe Framework of Reference (CEFR) for Languages. The CEFR is an international standard that provides an established range of levels of language proficiency necessary for various purposes. For example: B2 is widely seen as “the first level of proficiency of interest for office work” and additional studies have been conducted validating the use of the B2 level as necessary for academic success, for example (Carlsen, 2018). C1 is described as “effective operational proficiency”, and C2 as “practical mastery of the language as a non-native speaker” (North, 2007). As these levels have been tested and validated in a wide range of contexts and countries, they provide a robust measure of validity for Envoy.

Because these levels are descriptive and the needs for levels of proficiency differ among contexts, Envoy does not establish a “passing score,” but rather provides a description of the applicant’s proficiency so that the determination of sufficiency can be made by teachers or program administrators.

Test Items

Test items (i.e. questions and tasks) are written by native English writers with experience in both instruction and assessment of language learners. Item writers undergo training to become familiarized with the CEFR levels. Specifications for each test section were created that detailed the expectations for appropriate topics, item levels, and features of item quality. Items go through a process of peer-review for level, wording, and difficulty. Every item is field-tested with language learners of various proficiency levels.

All test items are subjected to Rasch analysis, which measures the relative levels of the questions. This ensures that applicants at lower levels of proficiency receive questions that are not challenging for those at higher levels, and applicants at higher levels receive questions that elicit the full range of their abilities. The use of Rasch analysis provides further validation of our levels (Karlin, 2018).

Construct Definition

Based on the purpose of the test described above, both foundational language knowledge (tested through discrete items), and communicative skills (tested through tasks related to reading, listening, writing, and speaking) were chosen for the test.

The CEFR was chosen as the primary framework of language proficiency of Envoy. Using this model allows for clear descriptions of proficiency levels to be provided to administrators and end users for use in instructional decisions. The CEFR is an internationally accepted proficiency scale that has as one of its main goals “facilitating transparency in testing and the comparability of certifications” ("Purposes of the CEFR," 2023).

Language knowledge was deemed important to assess because it helps instructors and curriculum developers determine what skills are most in need of direct instruction and remedial intervention to help learners. This knowledge is also generalizable to all domains of language use. The areas of language knowledge chosen – grammar, vocabulary, and cohesion – have been determined to be highly predictive of communicative language proficiency. Vocabulary knowledge is highly correlated with reading proficiency (e.g., Nation & Coady, 1988; Laufer, 1992, 1996; Wallace, 2007; Harkio & Pietila, 2016), knowledge of grammar with overall proficiency (Norris, 2005; Spinner, 2011), and understanding of cohesion with writing (Crossley, et al, 2016).

Communicative skills are the core of Envoy. The test assesses receptive (listening and reading) and productive (speaking and writing) language skills. The listening test evaluates comprehension through monologues and dialogues, simulating lectures and social interactions.

The reading test includes tasks that require processing informational and persuasive texts.

For the speaking test, test-takers engage with both concrete and abstract topics, with the complexity adjusted according to their proficiency level.

The writing test assesses two task types: email composition and essay writing. The answers are evaluated for both rhetorical and language skills, including topic development, style, organization, vocabulary, grammar, and mechanics.

All sections of the test are modular, allowing for sections, and in some cases parts of sections, to be administered in any combination relevant to the context.

The combination of testing language knowledge, reception, and production skills allows for a detailed diagnosis of the student's strengths and areas requiring further instruction or practice, providing crucial information for the design of instructional programs.

The content specifications of the questions were drawn from the Eaquals Core Inventory (British Council, 2010) which specifies the structures, vocabulary, and functions related to each level of the CEFR. Primarily, the Eaquals Core Inventory is designed for instructional purposes. The level of language that a student may encounter in a classroom may differ from what a student is able to produce independently (Kitao & Kitao, 1996). Research has shown that there are differences in the language a learner may know about, understand in context, and be able to produce (e.g., Macrory and Stone, 2000; Laufer, 2005). For example, a test taker may be able to select the correct word at level A2, understand it in a listening context at the B1 level, and produce it independently at the B2 level. For these reasons, grammar, vocabulary, and cohesion are assessed via both discrete items and open-ended items, and thereby give a full picture of a student's knowledge of language and ability to use that knowledge in language production.

Test Design Process

The test was designed in collaboration with experienced language teachers, content developers, and researchers. The process began with an ideation stage about which features would be most helpful for instructional programs, teachers, and language learners. Requirements that influenced the design of the test included:

- the need to predict proficiency across the four language skills: listening, speaking, reading, and writing
- clear and detailed scoring
- the time required to complete the test (under 90 minutes)
- fast turnaround of results (within two hours)

- suitability for the full range of levels from A1 to C2
- the ability to take the test on demand from any location
- the ability to select which skills will be tested.

Following the ideation process and specification of high-level requirements for the test, design features that would address such requirements were determined. Specifically, these were use of multi-stage adaptive technology and direct assessment of foundational language knowledge. The test design reflects the need to assess a students' level quickly and securely without prior indication of the level. Accordingly, a multi-stage adaptive test (MST) design was chosen. The first sections of the test measure foundational language knowledge via multiple choice questions. These questions indicate of what a student knows about language, and what requires further instruction. These sections also provide the start levels for the communicative skills portions of the test.

Item specifications for each section of the test were developed by language-testing professionals, based in part on the Eaquals Core Inventory. An existing item bank developed for an earlier test was used as a base. The items in the bank had already been field tested. A bank of test items was developed iteratively and piloted with a group of English language learners (n=150) in several different countries. After the initial pilot, levels of different sections were adjusted to more closely align to the mean levels and standard deviations of each section. Then another trial was conducted (n=20).

Multistage Adaptive Test Design

In order to provide an efficient testing experience, some sections (i.e. reading, grammar, cohesion, listening and reading) of the test are internally adaptive , and the score on each section provides the start level for a section that follows. In the four-skills version of the test, grammar is administered first, and provides the start level for the listening comprehension section. The vocabulary section provides the start level for the reading section, and the cohesion section provides the start level for the writing section. The listening comprehension score provides the level of the speaking questions.

The grammar test is divided into four levels of questions. The questions are divided among verbs, nominatives, adjectives and adjectival phrases, adverbs and adverbial phrases, modals, and prepositions, prepositional phrases, and phrasal verbs. Questions are aligned with the assumption that, if someone knows a more complex structure of a given type, then they probably know a simpler structure as well. As such, someone who knows the present perfect, for example, is presumed to know the simple present, and someone who can identify the correct use of a gerund phrase presumably can identify the correct plural form of a regular noun. Accordingly, the adaptivity of the grammar section is structured so that test takers start at the second level. As they answer questions correctly, they move up to higher levels of the same type of structure. As they make mistakes, they move down to lower levels. Within a level, they proceed until either three questions are answered correctly, or one is answered incorrectly. The testing level is determined by the level with the highest number of correct answers.

The vocabulary section has 6 different levels of vocabulary. Test takers start at the B1 level. Correct answers move test takers up a level, and incorrect answers move the test takers down a level. As students progress through the test, the movements among levels become smaller until the most likely level is obtained. The cohesion section functions similarly.

The listening section begins with a monologue at the end level of the grammar section (or one of the other multiple-choice sections if the test does not have a grammar section). If most questions are answered correctly, a higher-level monologue section is provided. If   questions are answered incorrectly, a lower-level monologue task is assigned. The final score is determined by the number of correct answers on the second task. A dialogue task is assigned based on the final level of the monologue section. The listening section proceeds as described above.

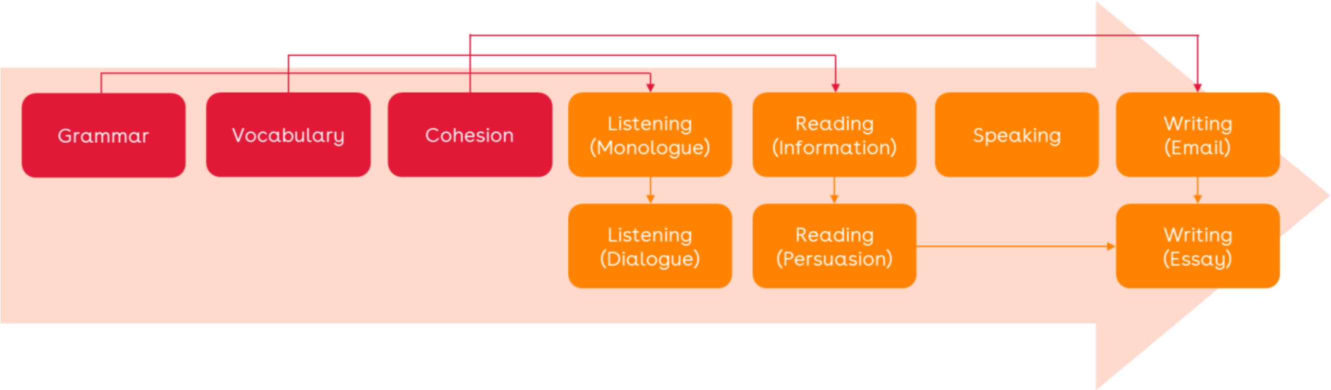
The reading section functions similarly to the listening section. There are two tasks, one for informational reading and one for persuasive texts. The start level is determined by the vocabulary multiple choice section, when it is present, or by one of the other multiple-choice sections if it isn't.

The speaking section is not internally adaptive. For the speaking test, there are two levels of questions covering domains of personal, academic, public, work, and other. The lower-level questions are more concrete and use lower-level vocabulary than the higher-level questions. The level of the questions that the students receive is not determinative of the final score;students can score at any level, regardless of the level of questions they receive.

The writing sections are also not internally adaptive. The writing test consists of email and essay tasks. As with the speaking section, the tasks are divided into two levels of tasks, but students can score at any level, regardless of the level of questions they receive. The start level of the email writing task is drawn from the cohesion section. The essay writing task is an integrated task, which requires an essay response to the reading task that received the highest number of correct answers.

Given the modularity of the test, if one of the sections is not included in the version, the start levels are drawn in the most logical way from the remaining sections. In the absence of grammar, vocabulary and then cohesion are used to determine the start level of the listening section. In the absence of vocabulary, grammar, and cohesion is used to determine the start level of the reading. In the absence of reading, cohesion is used to determine the start level of writing. In the absence of listening, grammar is used to determine the start level of speaking.

Figure 1. Envoy's multistage adaptive methodology



Test Content Development Process

Test Development Staff

All of the specifications for each section were developed by specialists with graduate level degrees in TESOL and linguistics who have extensive experience teaching adult English-language learners. The specifications include samples of questions of each type, along with directions about the factors to consider in writing each item. These factors include the appropriateness of content, difficulty level of items, range of vocabulary, level of grammatical structures, and length of items and input. There are also specifications about question quality, which include effectiveness of distractors and format of question stems.

Item writers are all experienced assessment professionals with many years of experience in teaching, writing, and rating tests.

Content Writing and Reviewing

All questions undergo a multiphase review process. Content reviews include checking for effectiveness of distractors, fairness of the item difficulty level of the item, and appropriateness of content. The review process included the following stages:

1. Item writing
2. Content review (appropriateness, level of items, effectiveness of question, effectiveness of distractors, etc.)
3. Revision
4. Content review
5. Editorial review
6. Final content review

Structure of the Test

The test is composed of adaptive multiple-choice questions, and open-ended questions requiring extended answers. Students are rated on knowledge of language, via discrete item assessment of grammar, vocabulary and cohesion, receptive knowledge of language via multiple choice questions about audio clips and written texts, and productive language ability via open ended speaking and writing tasks. Within speaking, users are evaluated on their vocabulary range, phonological control, fluency, cohesion, and grammatical range and accuracy. Within writing, users are evaluated on vocabulary range, grammatical range and accuracy, mechanics, topic development, organization, and style.

Each multiple-choice section of the test, including the listening and reading sections, is internally adaptive. Additionally, depending on the sections chosen, the final level of one section may be used as the start level of another. This practice allows a deeper focus on the strengths and areas for improvement in language proficiency while keeping the total administration time under 90 minutes when all sections are included.

The inclusion of sections that test language knowledge, receptive skills, and productive skills allows for a diagnosis of areas needing instruction and practice. Students may be able to recognize correct structures and define words without using them productively. Similarly, understanding individual lexical items is a prerequisite to comprehension skills, but is not identical to them. Testing various aspects of language proficiency allows instructors and program managers to identify areas most in need of improvement and to shape curriculum and instruction accordingly.

Grammar

The test's grammar section consists of 16-20 multiple choice questions testing a range of grammatical structures, including verb forms, articles, adjectives, adverbs, conditionals, modals, and nominatives. Each type of structure is tested by questions at a range of levels. These levels are based on the Eaquals Core Inventory, which was developed in alignment with the CEFR. The assumption is that, if a test taker can successfully answer a question on a higher-level structure, then they would be able to answer a question about a similar lower-level structure. For example, a student who can identify correct usage of a modal perfect--"I could have done it"--can most likely answer a question about the present perfect--("I have already eaten lunch." The grammar section is adaptive throughout.

Vocabulary

The vocabulary section consists of 16-20 multiple choice questions and is adaptive. Items were generated based on a list of leveled vocabulary extracted from speech samples tagged by expert raters. Once the items were generated, they were piloted on a sample of 150 students, and then levels were readjusted. Because the lists were generated from vocabulary production and the multiple-choice questions are testing vocabulary understanding, the levels of the questions were adjusted to a level below the levels of the tagged vocabulary. This increased the correlation between the vocabulary multiple choice and the vocabulary scores in the speaking section.

Cohesion

The cohesion section consists of 8-12 tasks, including 1 extended task that requires adding connectors to a paragraph. The task types include combining sentences, inserting connectors, and identifying errors. This section is adaptive. The cohesive structures were obtained from the analysis of writing samples of test takers at a range of levels with structures tagged by expert raters.

Listening Comprehension

This section contains 2-4 tasks, consisting of monologues and dialogues. Each task has 3-5 questions. The tasks are written and undergo multiple levels of review by expert teachers and raters. The section is adaptive.

Reading Comprehension

There are 2-4 reading tasks divided into informational and persuasive texts. Each text has 5 multiple choice questions that test understanding of vocabulary in context, understanding of main idea, details, and inference. The section is adaptive. Each text is analyzed for vocabulary level, readability, sentence structure, number of verbal elements per sentence and sentence length, and is written and reviewed by expert teachers and raters.

Speaking

The Speaking section consists of 5 open-ended questions. The choice to use open-ended questions allows the assessment to approximate authentic language use tasks, while still allowing a measure of control that allows for automated assessment.

The adaptive questions generate an estimate of the applicant's overall receptive language level as beginner, independent, or proficient. Based on this estimate, students are assigned questions designed to elicit their highest productive abilities lexically, textually, and grammatically. The open-ended questions increase in level of abstraction:questions at higher levels require more linguistic complexity and require the ability to take a stance on an issue;lower-level questions are connected to day-to-day and familiar topics, and require less grammatical and lexical sophistication, in alignment with the can-do descriptors of the CEFR.

The time allotted for answering each question is 2.5 minutes, although test takers are permitted to speak for less time if they choose. The speaking section is rated for vocabulary, grammatical range and accuracy, pronunciation, and fluency.

Vocabulary

Vocabulary range has been established as a necessary skill for success in academic and professional contexts. Research has established a threshold of a productive vocabulary of about 3000 words for university success (Nation, 1993, AB Manan et al, 2016). Vocabulary is rated via many features, including word frequency and word complexity.

Fluency

Features related to fluency, such as location and duration of pauses, have also been shown to contribute to judgments of a speaker's proficiency, as do grammar and vocabulary, to a lesser extent (Saito, et al, 2016).

Pronunciation

Pronunciation is frequently cited as a source of negative judgments of non-native speakers of English in both the scholarly and popular literature regarding international teaching assistants (e.g., Isaacs, 2008) and business professionals (Executive Education, 2013). While it is necessary to acknowledge that many of these judgments may reflect biases or judgments of non-linguistic proficiencies, presenting the pronunciation scores together with the rest of the applicant's language proficiencies allows a determination to be made of the speaker's intelligibility, to what extent the accent may or may not impede communication, and to what degree the pronunciation reflects a more general linguistic proficiency.

Grammar

According to most linguists and language teachers today, "the primary goal of language learning today is to foster communicative competence, or the ability to communicate effectively and spontaneously in real-life settings" (Purpura, 2004). Envoy distinguishes between judgments of grammatical accuracy and the ability to produce grammatically accurate speech. The first measure, along with questions about vocabulary and communicative knowledge, is used to gain a general proficiency level (beginner, intermediate, or advanced). The ability to use grammar effectively in speech is judged through assessment of the open-ended questions.

Writing

The writing section consists of two tasks: writing an email and writing an essay. The email is a response to a prompt of an academic or work-related situation. The essay is an integrated task that requires a response to a text, the type of task that would be expected of students in the final years of high school and throughout university. In tests that include a reading section, the text is one of the texts they have read in the reading comprehension test. The writing tasks are rated for vocabulary, grammatical range and accuracy, mechanics (spelling, punctuation, paragraphing), topic development, organization, and style (register, genre conventions, appropriateness).

Scoring

All sections of the test are scored according to the CEFR, and the CEFR level scores are presented in the score report. For the multiple-choice sections, the score is determined by the level with the largest number of correct items. Because the test adapts, when an item is answered incorrectly, a lower-level question is presented, and when an item is answered correctly, a higher-level item is presented. Accordingly, the level with the highest number of correct items can be assumed to represent the student's correct level.

Open-ended questions are scored according to CEFR-aligned rubrics. The questions are scored by the AI Rater. The AI Rater consists of algorithms that were trained on data scored by experienced CEFR exam raters in a blind-rating process, which uses two raters to independently score the speech and writing samples, and a third independent rater used to resolve discrepancies. After AI scoring of the tests, human raters provide an additional layer of quality control, with an additional check used in the case of discrepancies.

All raters have prior experience rating other CEFR-aligned tests. They go through internal training before rating Envoy tests. Raters are monitored for quality on an ongoing basis, with more training provided for those whose level of agreement is deemed too low.

The AI Rater is continually validated against the scores of human raters. A golden data set, consisting of six independent ratings of 1000 tests initially, with additional sets added when new populations begin using the test is used to ensure precision of the AI Rater. The ability to predict the average of three human ratings by another three human ratings is compared to the AI rater's ability to predict the first average. The AI rater is considered sufficiently accurate to replace an individual rater when the ability to predict the human score is within 2 percentage points of the human accuracy levels.

Test Administration, Security and Privacy

The Envoy test combines human expertise with the power of AI to ensure that results of the test are fair and trustworthy. Each test is recorded via the computer, camera, and microphone, and when suspicious behavior is detected, the results are reviewed by multiple rounds of trained professionals.

Test Design

The adaptive format of the test, along with tracking of administered questions ensures that candidates receive different questions each time they take the test.

Pre-test

A link to enter the test is sent to the email address through which the administrator has registered the test-taker. Only this email address can be used to enter the test.

Once the test taker has entered the testing environment, a photo is taken of the test taker which is presented on the report. This allows the test administrator to confirm that the test taker is the intended candidate.

Browser extensions are blocked throughout the test, ensuring that external applications can not be used during the test.

During the test

Various data points are collected throughout the test. AI is used to detect anomalous patterns and flag exams for human review.

If the test-taker leaves the window of the test repeatedly or for an extended period of time, the test is automatically invalidated.

After the test

Test answers are analyzed for irregularities, including use of external sources, and data collected throughout the test is analyzed and reviewed when necessary.

Samples of the test-taker's speech and writing are presented on the score report to allow additional checks on the part of the test administrator.

Envoy fully complies with the EU's General Data Protection Regulation (GDPR) to safeguard test takers' personal data and privacy. As a data processor, Envoy follows precise guidelines and protocols as dictated by its users and data controllers to handle personal data securely throughout all assessment activities. This involves implementing rigorous access controls, allowing only essential staff to access data based on least privilege principles. Additionally, the test employs GDPR-aligned security measures such as encryption and pseudonymization to process data lawfully, fairly, and transparently.

Central to Envoy's privacy policy is limiting data collection to only what is required for delivering tests, respecting data minimization obligations. Envoy exercises full transparency in use of data making clear it analyzes recordings to improve algorithms and services while anonymizing data. Furthermore, Envoy details test takers' rights under the GDPR, pledging to assist customers in addressing requests like data access, rectification, and erasure. By contractually binding itself as a processor and maintaining liability over any sub-processors appointed, the test aims to provide robust privacy protection aligned with the GDPR.

To ensure data security, the test implements appropriate technical and organizational controls including potential use of encryption, firewalls, and secure data centers. Role-based access control, network segmentation, and periodic auditing of systems also protect personal data. In case of any data breaches, Envoy follows a defined procedure for promptly notifying impacted controllers.

Our privacy policy informs users of their rights by stating that data subjects can contact the customer controller (administrator) to exercise rights like access, rectification, and erasure of personal data.

Research

After full rollout of the test, further research will be conducted on repeated administrations, item performance, section performance, and AI and human rater reliability.

References

ab Manan, Nor & Azizan, Noraziah & Fatima Wahida Mohs Nasir, Nur. (2017). "Receptive and Productive Vocabulary Level of Diploma Students from a Public University in Malaysia." *Journal of Applied Environmental and Biological Sciences*. 7. 53-59.

British Council/EAQUALS (European Association for Quality Language Services). (2010) *British Council – EAQUALS Core Inventory for General English*. https://www.eaquals.org/resources/the-core-inventory-for-general-english/

Carlsen, C. H. (2018). "The Adequacy of the B2 Level as University Entrance Requirement." *Language Assessment Quarterly*, 15(1), 75–89. doi: 10.1080/15434303.2017.1405962

Executive Education. (2013, December 7). "The Glass Ceiling Facing Nonnative English Speakers". Retrieved from https://knowledge.wharton.upenn.edu/article/glass-ceiling-facing-nonnative-english-speakers/

Hulstijn, J. H. (2002). "Towards a Unified Account of the Representation, Processing and Acquisition of Second Language Knowledge." *Second Language Research*, 18, 193–223.

Isaacs, T. (2008). "Towards Defining a Valid Assessment Criterion of Pronunciation Proficiency in Non-Native English-Speaking Graduate Students." *Canadian Modern Language Review*, 64(4), 555–580. doi: 10.3138/cmlr.64.4.555

Jaroszek, M. (2011). "The Development of Conjunction Use in Advanced L2 Speech. Studies in Second Language Learning and Teaching," 1(4), 533–553. Retrieved from https://search-ebscohost-com.ezproxy.snhu.edu/login.aspx?direct=true&db=eric&AN=EJ1136573&site=eds-live&scope=site

Karlin, O., & Karlin, S. (2018). "Making Better Tests with the Rasch Measurement Model." *InSight: A Journal of Scholarly Teaching*, 13, 76–100. Retrieved from https://files.eric.ed.gov/fulltext/EJ1184946.pdf

Macrory, G., & Stone, V. (2000). "Pupil progress in the acquisition of the perfect tense in French: The relationship between knowledge and use." *Language Teaching Research*, 4(1), 55-82.

Nation, I.S.P. (1993) "Vocabulary size, growth and use." *The Bilingual Lexicon*. R. Schreuder and B. Weltens (eds.), Amsterdam/Philadelphia: John Benjamins: 115-134.

North, B. (2007, February 6). *Common European Framework of Reference for Languages (CEFR)*. Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages/documents

Purposes of the CEFR - Common European Framework of Reference for Languages (CEFR) - www.coe.int. (n.d.). Retrieved from Common European Framework of Reference for Languages (CEFR) website: https://www.coe.int/en/web/common-european-framework-reference-languages/uses-and-objectives#:~:text=The%20CEFR%20is%20intended%20to,of%20the%20Council%20of%20Europe.

Purpura, J. (2004). "Differing notions of 'grammar' for assessment." *Assessing Grammar* (Cambridge Language Assessment, pp. 1-23). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511733086.002

Sato, Takanori. (2013). "The Influential Features on Linguistic Laypersons' Evaluative Judgments of Second Language Oral Communication Ability." *JLTA Journal* 16. 107-126.

10.20622/jltajournal.16.0_107.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). “Second Language Speech Production: Investigating Linguistic Correlates of Comprehensibility and Accentedness for Learners at Different Ability levels.” *Applied Psycholinguistics,* 37(2), 217-240. doi: http://dx.doi.org.ezproxy.snhu.edu/10.1017/S0142716414000502