# AI-EnglishPro Reliability & Validity Manual

***Prepared By:***
iMocha, Inc.
[support@imocha.io]

# Table of Content

# Introduction

iMocha's AI-EnglishPro Speaking and Writing assessments are powered by NLP Models and Generative AI technology. It is Business English assessment instrument tool designed to objectively measure proficiency of a person in speaking and writing the English in a business context. This assessment tool is intended for talent leaders, both recruiters and talent managers, to evaluate business English proficiency with minimum time and efforts. The test can be taken on a mobile or laptop device, and since it is online, it can be taken from any location. It eliminates the need for human expertise to examine the skills.

The automatic AI-driven report generation mechanism enables objective, bias-free, instant and reliable results. The score rating include CEFR standards, measures important parameters such as vocabulary, fluency, grammar, etc, in addition to a completely AI-enabled contextual check. The talent analytics provided by the AI-EnglishPro product is a reliable source of intelligence that can be used to make smarter decisions during hiring drives, talent management and professional certifications

# 1. Test Description

## 1.1 Test Design

The AI-EnglishPro Speaking and Writing assessments are compatible with mobile and laptop devices and can be taken from any location at any time. The test consists of two question types: Speaking – Open Questions and Writing – Business Email Writing Questions. Both these sections record responses which are analyzed automatically by the backend AI engine and NLP models to generate relevant scores. The tool is based on CEFR ratings and generates talent analytics on a range of parameters that include iMocha's overall proficiency score, oral fluency, Grammar including whitespace and typo mistakes, Vocabulary, Word count, CEFR level of Speaking and Writing, CEFR Vocabulary Categorisation, Readability and Linguistic Diversity. An important measure is the degree of Digression from business context in the spoken and written response. This is an AI-generated measure which identifies how aligned a response is to a given business context.

The AI-EnglishPro tool automatically analyses talent's response and generates reports within few minutes. The scores can be downloaded as pdf. Ranking of test takers is also available on the platform to help in quick filtering. The questions in the test can be customised to fit a given industry, use case or business scenario and the AI-system as well as NLP models generate reliable reports on customized content as well.
The tests can be combined with technical assessments as well and reports are generated for both tests.

## 1.2 Test Administration

Administration of the AI-EnglishPro assessments takes 15-20 minutes on a mobile device or computer. The tool allows test taker to click on Submit button once they are done with their response. Clicking Submit ensures response is recorded and ready for AI analysis and leads test taker to next section of their respective test.

### 1.2.1 Mobile App Administration

For mobile administration, the test can be taken on any browser on the Android system and works well with Safari on the iOS devices. The test platform can be accessed using a test link shared prior to test date. The microphone settings need to be enabled on the browser for successful test attempts on Speaking questions. The mobile device must have Internet connection and if the test is proctored, the

camera settings must be enabled. The instructions for each of the sections are displayed on the screen before the timer begins. Test taker can record their response by speaking on the microphone in English. The test taker gets two attempts to record their response in the given time limit. When a test is finished, the test taker must click on "SUBMIT" button in order not to lose the response, for both Speaking and Writing sections.

### 1.2.2 Computer Administration

For computer administration, the test can be taken using the test link which will open in both Chrome and Safari browsers. The microphone settings need to be enabled for the browser for successful test attempts on Speaking questions. The laptop must have Internet connection and if the test is proctored, the camera settings must be enabled. The instructions for each of the sections are displayed on the computer screen before the timer begins. Test taker can record their response by speaking on the microphone in English. The test taker gets two attempts to record their response in the Speaking section, in the given time limit. When a test is finished, the test taker must click on "SUBMIT" button in order not to lose the response.

## 1.3 Test Format

The test can be accessed using the link which is shared prior to test via email or any other channel. The link is valid for a limited time period as set by the test administrator. Test taker has to access the link before it expires. The link leads the candidate/employee into the test platform window which requires login. After login, a set of instructions are shown. The microphone and/or camera will need to be enables for the system based on test requirements. On starting the test, the respective sections appear with the timer set as per total time duration for each section.

iMocha has two question types; AI-evaluation: Speaking and AI-evaluation: Writing. These questions can be added as a separate section to any technical assessments as well. The following sections give a brief description of the types of questions in an AI-EnglishPro test:

### 1.3.1 PART A: Speaking

This section consists of Open Questions based on real life business scenarios.

A scenario based open question appears on the screen. Candidate/employee has to think about their response, start the recording and record their response. The

platform allows for retakes as per the administrator's needs. After recording, test taker must click on SUBMIT button for the response to be considered for scoring.

### 1.3.2 PART B: Writing

This section consists of a business Email writing question based on real life business scenarios.

On entering this section, a business scenario is described on which the candidate/employee has to draft a professional email to the indicated recipients. After completing the email, candidate/employee must SUBMIT the response.

In iMocha's AI-Speaking and AI-Writing test, the test taker enjoys the entire creative space, without any restrictions, to develop answers. These response are analyzed by the AI-enabled context check which measure how aligned they are to the scenario depicted in the question.

## 1.4 Number of Items

In the AI-EnglishPro AI-powered question types, there are two sections each for AI-Speaking and AI-Writing. In a standard test, these sections contain one question each which is exhaustive in itself to test business English skills in the given time limit. On iMocha's platform, these can be customized to include more than one questions to be attempted in the desired time limit. The exact number of items in each test may change based on the organization's needs. The scores generated for the responses to standard test and customized test are equally reliable and valid.

## 1.5 Test Construct

AI-Speaking: This is a video type question where the candidate/test taker is provided with a situation-based question, and they need to submit their live recorded answer. When the candidate/test taker submits their answer, the recording is auto-transcribed using an accurate speech-to-text transcription. Our NLP models evaluate their submitted response based on these parameters – Oral Fluency, Vocabulary and CEFR level ratings.

AI-Writing: This is a writing type of question where the candidate/test taker is provided with a situation-based question, and they need to write an email addressing the issue specified in the question. When the candidate/test taker submits the answer, our NLP models evaluate the answer based on these parameters – CEFR level, Grammar, Readability, Vocabulary, Linguistic Diversity and Word count.

# 2. Content Design and Development

The AI-EnglishPro Speaking and Writing tests have been designed to focus exclusively on business use cases, emphasizing on the importance of professional and business English for organizational operations, growth, and revenue. It does not consider academic or usual everyday English style. The test focuses on the ability of both native as well as non-native speakers to speak and write in English that is critical for businesses, at a global level. The responses are digitally stored and analyzed by the AI-capabilities of Natural Language Processing models and AI-powered context check. Responses can be analyzed automatically to generate reliable metrics and data that will guide talent decisions for businesses. These include fluency, vocabulary, grammar, CEFR ratings as the most basic parameters.

Sub-scores such as linguistic diversity and readability in the Writing Section and Oral fluency in the Speaking Section have been designed to refine talent decisions.

## 2.1 Vocabulary Selection

The vocabulary used in the test is based on the 14-billion-word corpus, i.e., the iWEB corpus which is globally accepted and renowned. The language structures used in the test are derived from everyday business conversations and exchanges. This includes extensive use of professional language, less use of slangs and common everyday English outside businesses and also takes into consideration the ethical construct when conducting professional conversations, whether while speaking or writing.
The scores generated take into consideration the high and low frequency vocabulary, showing high score when a lower frequency word has been used by the test taker and lower score for the use of more mundane and high frequency words.

In the AI-Writing Section, the vocabulary can be further broken down into CEFR levels which is depicted as Vocabulary Categorisation in the report. Here, one can measure the percentage of words used by the candidate that belong to a particular CEFR level category. For Example, the report provides answer to *What is the percentage of words used by test taker fall in C1 level of CEFR?*

## 2.2 Item Development

The items in the AI-EnglishPro assessments were drafted by internationally qualified subject matter experts who belong to the Mochaworks community at

iMocha. The language structures in the test items reflect those of real life business scenarios and are constructed in the professional tone. Based on the intelligence from subject matter experts and study conducted with globally qualified IO psychologist, the items of the test were drafted. The idea was to ensure that the content is aligned to measure the intended skills for an industrial job role. Over three thousand sample items were considered in the process which were then reviewed internally by a team of test developers to ensure the conformity with global business standards. Compliance with the vocabulary specification and conformity with usage of business English globally was checked.

## 2.3 Score Reporting

### 2.3.1 Scores Weightage and Use

The AI-Speaking and AI-Writing reports comprises of an overall score and sub-scores as such – Oral Fluency, Vocabulary, Grammar, and Word Count. The corresponding CEFR level (Common European Framework of Reference for Languages) and the Readability score of Flesch Reading ease are also displayed.

For AI-Speaking, the CEFR level, Oral Fluency, and Vocabulary usage contribute towards the overall speaking score –

| Sub-scores | Weightage |
|---|---|
| CEFR | 25% |
| Oral Fluency | 50% |
| Vocabulary | 25% |

- Oral Fluency: We have benchmarked the average words-per-minute count across thousands of candidates. Based on the words-per-minute spoken by each candidate, they are scored on their overall oral fluency.
- Vocabulary score: Using a corpus of more than 14 billion+ words, and their overall words usage frequency, our NLP models generate a weighted score for every unique word uttered by the candidate in their recorded answer.
- CEFR Level: Using the English Vocabulary Profiling, we determine the candidates overall speaking proficiency in terms of the CEFR levels.

For AI-Writing, the CEFR level, Grammar, Readability, Vocabulary, and Word Count contribute towards the overall writing score –

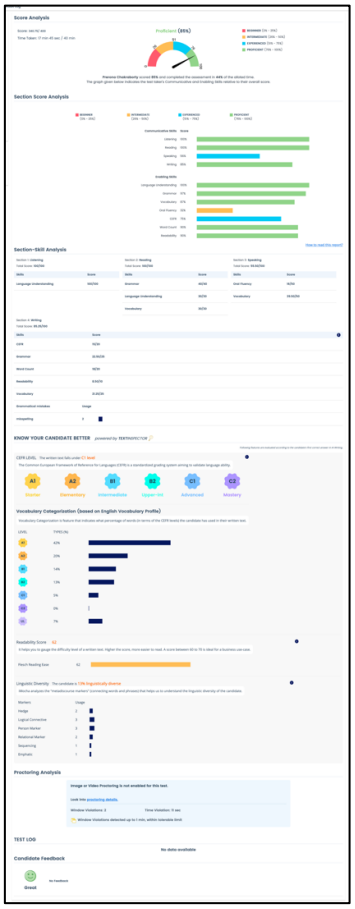
| Sub-scores | Weightage |
|---|---|

| | |
|---|---|
| CEFR | 20% |
| Grammar | 25% |
| Readability | 10% |
| Vocabulary | 25% |
| Word Count | 20% |

- CEFR level: Using the English Vocabulary Profiling, we determine the candidates overall writing proficiency in terms of the CEFR levels.
- Grammar score: Our NLP models identify the type of grammatical mistakes made by the candidate in their answer. Depending on the frequency of mistakes against the overall answer, a score is provided to determine the grammatical accuracy of the submitted answer.
- Readability score: We leverage the Flesh Reading Ease algorithm to determine the comprehensibility of the submitted answer. A complex and difficult answer would get a low readability score and vice versa.
- Vocabulary score: Using a corpus of more than 14 billion+ words, and their overall words usage frequency, our NLP models generate a weighted score for every unique word used by the candidate in their submitted answer.
- Word Count: Our NLP models analyze the overall answer submitted based on the total number of words used. The score is evaluated after excluding the usage of “stop words” and focusing on the important or actual words used throughout the answer.

Once a candidate has completed a test, the NLP engines analyze the spoken and written answers and post the scores to the test administration platform. Test administrators can choose to make scores available to test takers. Scores from the Versant English Test have been used by commercial and business organizations. The AI-Speaking and AI-Writing assessments enable the talent managers in making valid decisions about business English skills of individuals. These scores may also be used effectively in evaluating whether an individual’s level of English is sufficient to perform certain tasks or functions effectively. Talent managers and recruiters may wish to base their selection of an appropriate cut score based on their internal research. iMocha can provide benchmarking and further assistance in establishing cut scores.

### 2.3.2 Score Interpretation



The Score Analysis section provides an overview of the candidate's performance, with the Section Score Analysis dividing the scoring into two main components: Communicative Skills and Enabling Skills. A bar graph visually represents the candidate's percentage score for each skill, with color-coding corresponding to performance categories.

The Section-Skill Analysis delves into the candidate's performance within each section, focusing on various skills. In the writing section, the analysis highlights Grammatical mistakes, showcasing the type and frequency of errors made by the test taker.

The "Know your Candidate Better" section offers insights into the candidate's performance, comparing it to global benchmarks using the CEFR levels. Additionally, it utilizes the Flesch Reading ease metric to assess the readability of the candidate's writing abilities. iMocha also employs English Vocabulary Profiling to demonstrate the test taker's use of CEFR level words and linguistic proficiency.

# 3. Validation and Reliability

## 3.1 Study and Sample Size

A total of 2000 reports were reserved from a comprehensive dataset containing performance data from both native and non-native English speakers for the purpose of conducting validation analyses on the AI-Speaking and AI-Writing scoring mechanisms. The validation process encompassed two key aspects: Validity Testing and Reliability Assessment.

**Validity Testing:** Validation analysis was employed to ascertain the accuracy of the AI model's measurement instrument or score in assessing the intended construct. This entailed evaluating the instrument's discriminant validity, particularly in distinguishing between native and non-native candidates. Discriminant validity gauges the degree to which different constructs or scales are distinctly separate.

**Reliability Assessment:** The assessment included an interpretation of Cronbach's Alpha values. Higher values were indicative of stronger internal consistency among the assessment items. Furthermore, the Intraclass Correlation Coefficient (ICC) was calculated to gauge the level of agreement between AI-generated scores and human evaluator scores. This was accomplished through the implementation of a two-way random-effects model for absolute agreement.

**Internal Consistency:** Cronbach's Alpha coefficients were computed for both AI-Speaking and AI-Writing assessments to assess the internal consistency of items within each assessment.

**Correlation Analysis:** A comparison was made between the AI-generated scores and external measures of proficiency, such as validated language proficiency tests. This comparison was facilitated by calculating correlation coefficients (Pearson's r) to determine the strength of alignment between the two sets of scores.

In summary, the findings from the content and construct validity assessments were discussed, with particular attention to how well the assessment content aligned with the intended skills and proficiency frameworks.

## 3.2 Reliability

### 3.2.1 AI-Speaking (Native & Non-Native)

The Cronbach's Alpha coefficient for the total scale (AI-Speaking assessment) is provided in the table below. Typically, a Cronbach's Alpha value of 0.70 or higher is considered acceptable for most educational and psychological assessments. It reflects good internal consistency.

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| .533 | .807 | 10 |

Based on this analysis, the AI-Speaking assessment scores demonstrate good internal consistency as indicated by positive correlation coefficients and potentially satisfactory Cronbach's Alpha values. The individual items have positive correlation coefficients with the Vocabulary score and Oral Fluency Score, indicating that they are contributing to the overall assessment construct.

### 3.2.2 AI-Writing (Native & Non-Native)

**AI-Writing (Native & Non-Native)**

The Cronbach's Alpha coefficient for the total scale (AI-Writing assessment) is provided in the table below. Typically, a Cronbach's Alpha value of 0.70 or higher is considered acceptable for most educational and psychological assessments. It reflects good internal consistency.

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| .602 | .702 | 9 |

Based on this analysis, the AI-Writing assessment scores demonstrate good internal consistency as indicated by positive correlation coefficients and potentially satisfactory Cronbach's Alpha values. The individual items have positive correlation coefficients with the Vocabulary, Grammar, and Word Count scores, indicating that they are contributing to the overall assessment construct.

In summary, the reliability analysis output for Native and Non-Native Writing Assessment scores assessed by the AI assessment model provides insights into the internal consistency and reliability of the scores. High Cronbach's alpha values indicate consistent and reliable measurement of writing proficiency, while lower values might necessitate further investigation and potential improvements to the assessment model.

## 3.3 Validity

### 3.3.1 Vocabulary Scores

Wilks' Lambda is used to assess the separation between groups. A lower value implies better separation, signifying that the native and non-native groups have clearer distinctions in terms of Vocabulary Scores. If the p-value is low (typically below 0.05), it allows us to reject the null hypothesis, indicating a substantial difference in Vocabulary Scores between native and non-native individuals. In Vocabulary scores, the Wilks' Lambda value, which measures separation, was found to be statistically larger (p = 0.74, greater than 0.05), indicating a weak distinction between the groups.

| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
|---|---|---|---|---|
| 1 | .740 | 643.285 | 7 | .000 |

### 3.3.2 Grammar Scores

We conducted an analysis to examine the differences in Grammar Scores between native and non-native individuals using Discriminant Analysis. Our findings provide compelling evidence suggesting that there is no noteworthy difference between the two groups. The Wilks' Lambda value, which measures separation, was found to be statistically larger ($p = 0.943$, greater than 0.05), indicating a weak distinction between the groups in terms of Grammar Scores.

| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
|---|---|---|---|---|
| 1 | .943 | 125.964 | 7 | .000 |

The canonical discriminant functions, along with their eigenvalues and coefficients, offered valuable insights into how these functions contribute to distinguishing the groups. Our classification results indicated a high accuracy in predicting group membership based on Grammar Scores.

Taking into account both statistical and practical significance, we can confidently conclude that Grammar Scores play a significant role in distinguishing between native and non-native individuals. This study underscores the importance of Grammar Scores as a discriminating factor and highlights potential implications for language learning and assessment.

Our investigation into the disparities in Vocabulary Scores between native and non-native individuals using Discriminant Analysis has yielded insightful outcomes. Our analysis demonstrates that there is no statistically significant substantial distinction between the two groups.

Hence, we can confidently affirm that there is no significant difference in Vocabulary Scores between native and non-native individuals. Both from a statistical and practical standpoint, it is evident that Vocabulary Scores and Grammar Scores do not serve as distinguishing factors between native and non-native individuals. This finding has implications for the assessment methods employed, suggesting that the AI assessment model used does not effectively differentiate between native and non-native individuals. The insights derived from this study contribute to our understanding of the role of vocabulary in discerning language proficiency levels.

# 4. Conclusion

Following conclusions can be made based on the data studies:

- The AI-EnglishPro assessment produces reliable business English skill estimates.
- Sub scores of the test are reasonably distinct and offer diagnostics that are useful for corporations and business globally.

To assure the defensibility of employee selection procedures, employers in the U.S. follow the Equal Employment Opportunity Commission's (EEOC's) Uniform Guidelines for Employee Selection Procedures. These guidelines state that employee selection procedures must be reliable and valid. The above information provides evidence of the reliability, validity and legal defensibility of the AI-EnglishPro Speaking and Writing assessments is in conformance with the prescriptions of the EEOC's Uniform Guidelines.

Brewing A *Skills-First* Planet

Call us **+1-408-759-4457** or write to us at **support@imocha.io**

**About iMocha**

iMocha is an AI-powered Skills Intelligence Cloud™ that helps enterprises to build a Skills-First and data-driven ecosystem of hiring, upskilling, and managing talent - at scale, for any job role, and any industry. More than 500 organizations in 70+ countries are using iMocha's solution for taking a Skills-First approach to accelerated hiring, objective learning and development programs and to manage talent from candidate to alumni. Enterprises from IT/ITeS, Telecom, Banking, Financial Services and Insurance are using iMocha's Skills Intelligence platform to make data driven talent decisions. The platform leverages patented technologies and includes innovative features to build employee skills profile, organizational skills inventory, skills taxonomy and skills ontology, skill benchmarking and skills analytics that helps to hire, develop and manage talent by taking a Skills-First approach..

To know more, visit us at www.imocha.io.