English for Professionals Exam

Test Description and Validation Summary
Ver 1.0 - March 10, 2013

Table of Contents

Section I – Test Description

1. Introduction

1.1 Overview

Pearson's English for Professionals Exam (or E^Pro™) is a four-skill, computer-based assessment instrument which is designed to measure how well a person can handle workplace English. E^Pro is intended for adults 18 years of age and older and takes about 90 minutes to complete. The candidate registers online for the test via a designated website and the test can then be taken only at VUE test centers or VUE-approved test providers. The test is administered and scored entirely by computer without needing a human examiner or rater; however, a test proctor is provided by the test center. Because the test items are delivered and scored by an automated testing system, it allows for standardized item presentation as well as immediate and objective results that are reliable and correspond well with traditional measures of English language proficiency.

E^Pro is comprised of fifteen item types (Parts A through O). The majority of the item types integrate more than one language skill in performing the tasks, as in Table 1.

Table 1. Test Structure and Required Language Skills

Part	Task	Required Language Skills
A	Picture Description	Writing
B	Sentence Completion	Reading, Writing
C	Passage Reconstruction	Reading, Writing
D	Summary Writing	Reading, Writing
E	Reading Comprehension	Reading
F	Email writing	Reading, Writing
G	Dictation	Listening, Writing
H	Response Selection	Listening
I	Conversations	Listening, Speaking
J	Passage Comprehension	Listening, Speaking
K	Passage Reading	Reading, Speaking
L	Repeat	Listening, Speaking
M	Sentence Builds	Listening, Speaking
N	Speaking Situations	Listening, Speaking
O	Story Retelling	Listening Speaking

All items in E^Pro elicit responses from the candidate that are analyzed automatically. These item types provide multiple, independent measures that underlie facility in English, including sentence comprehension and construction, passive and active vocabulary use, and appropriateness and accuracy in speaking and writing.

The E^Pro score report is comprised of an Overall score, four skill scores (Speaking, Listening, Reading, and Writing) and eight analytic subscores with four coming from Speaking (Sentence Mastery, Vocabulary, Pronunciation, and Fluency) and four coming from Writing (Grammar, Word Choice, Organization, and Voice & Tone). Because more than one item type typically contributes to multiple subscores, the use of multiple item types strengthens score reliability.

The Overall score is a weighted average of the four skill scores, and the skill scores themselves are made up of weighted averages of the analytic subscores.

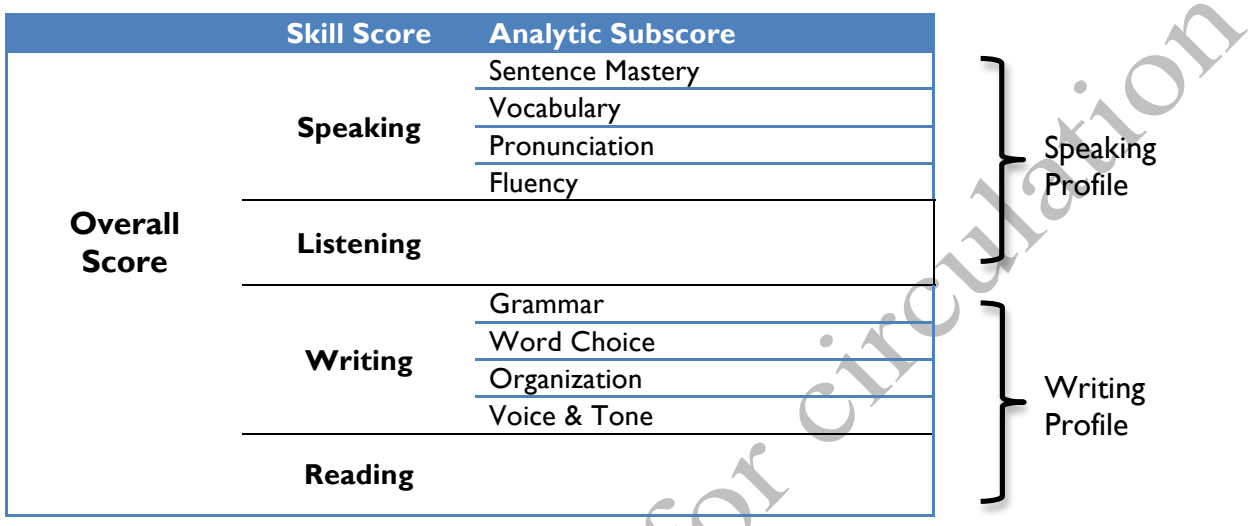

Figure 1. Overview of different scores reported in E^Pro score report

Furthermore, the four speaking analytic subscores and Listening skill score are combined to provide a broader picture of the candidate's spoken English skills, called Speaking Profile. Similarly, the four writing analytic subscores and Reading skill score are combined to reflect the candidate's broader written English skills, i.e., Writing Profile. For each of these two skill profiles, an overall performance description is provided. Together, these scores provide a comprehensive profile of the candidate's facility in English in everyday and workplace contexts.

The Versant testing system automatically analyzes the candidate's responses and posts scores to the VUE Credential Manager website within 5 business days of completing the test. Test administrators and score users can view and print out test results.

1.2 Purpose of the Test

E^Pro is a four-skill test that is designed to measure *facility* in spoken and written English in the workplace context, which is a key element in successful business communication in spoken and written English. Facility is defined as *the ability to understand spoken or written English on everyday and workplace topics and respond appropriately at a functional pace*. The test scores provide reliable, objective, and useful information about the level of English language proficiency of English learners. E^Pro test scores are primarily intended for use by businesses and government agencies where assessment of English language proficiency is an important part of recruitment, training, and advancement decisions. The test scores may also be used for monitoring progress as well as measuring instructional outcomes. Furthermore, E^Pro's analytic subscores provide information about the candidate's strengths and weaknesses, which may support instruction and individual learning.

The E^Pro score scale covers a wide range of abilities in spoken and written English communication. In most cases, score users must decide which E^Pro overall score and/or skill scores should constitute a minimum requirement in a particular context (i.e., a cut score). Score users may wish to base their selection of an appropriate cut score on their own localized research. Pearson can provide assistance in establishing cut scores.

In summary, Pearson endorse the use of E^Pro scores for making decisions related to test-takers' spoken and written English proficiency, provided score users have reliable evidence confirming the identity of the individuals at the time of test administration. Supplemental assessments would be required, however, to evaluate test-taker's academic or professional competencies.

2. Test Description

2.1 Workplace Emphasis

E^Pro is designed to measure the candidate's ability to understand and use English in workplace contexts. The test does not target language use in one specific industry (e.g., banking, accounting, travel, health care) or job category (e.g., shop clerks, accountant, tour guide, nurse) because assessing the candidate's English ability in such specific domains requires both English ability and content knowledge, such as subject matter knowledge or job-specific terminology. Rather, E^Pro is intended to assess how well and how efficiently the candidate can process English on general topics such as scheduling, commuting, and training, which are commonly found in the workplace regardless of industry or job category.

2.2 Test Administration

E^Pro is a computer-based test. The test can be taken only at VUE test centers or VUE-approved test providers by using computers that have already installed specific Pearson's test software. The candidate is fitted with a microphone headset. The test software prompts the candidate to adjust the volume and calibrate the microphone before the test begins. The computer and microphone headset are provided by VUE test centers.

It is best practice to provide the candidate with the test-taking tutorial before the actual testing begins so that the candidate can become familiar with the test format. (Please contact Pearson for this tutorial.)

The instructions for each section are spoken by a recorded examiner voice and are also displayed on the computer screen. Candidates interact with the test system in English, using the following response modalities: typing using a keyboard, clicking using a mouse, and speaking into the microphone. When the test is finished, the candidate may leave the testing station.

The candidate has a set amount of time to respond to each item. A timer can be seen in the upper right corner of the computer screen during all tasks except for the speaking tasks. The delivery of the recorded item prompts for speaking items is interactive – the system detects when the candidate has finished responding to one item and then presents the next item. For written items, if candidates finish before the allotted time has run out, they can click a button labeled "Next" to move on to the next item. If candidates do not finish a response in the allotted time, their work is saved automatically and the next item begins.

2.3 Number of Items

E^Pro is a 90-minute, four-skill English test. During each test administration, a total of 107 items are presented to each candidate in the fifteen separate sections, Parts A through O. In each section, the items are drawn from a large item pool. For example, each candidate is presented with eighteen Sentence Completion items selected quasi-randomly from the pool, so that most items will be different from one test administration to the next. Items are selected for each test form based on, among other things, the item's level of difficulty and its form and content in relation to other selected items. Table 2 shows the number of items presented in each section.

Table 2. Number of items presented per section

Part	Task	Presented
A	Picture Description	1
B	Sentence Completion	18
C	Passage Reconstruction	3
D	Summary Writing	1
E	Reading Comprehension	12 (6 passages, 2 questions each)
F	Email writing	2
G	Dictation	14
H	Response Selection	8
I	Conversations	10
J	Passage Comprehension	6 (2 passages, 3 questions each)
K	Passage Reading	2
L	Repeat	14
M	Sentence Builds	10
N	Speaking Situations	3
O	Story Retelling	3
	Total	107

2.4 Test Format

During the test administration, each task is introduced with instructions and, where informative, an example. The instructions and example for the tasks are spoken in an examiner voice and are also displayed on the computer screen. Test items with audio present recordings that are spoken by a variety of native English speakers including American, British, and Australian dialects, as well as highly proficient non-native speakers. Voices of these test items are distinct from the examiner voice.

The following subsections provide brief descriptions of the item types and the abilities required to respond to the items in each of the fifteen parts of E^Pro.

Part A: Picture Description

In the Picture Description task, candidates see a picture and are asked to describe what is happening in it.

Example:

> Picture Description. Look at the picture below. Write a description of the picture in English. You have two minutes. Write as much as you can. Use complete sentences.
>
> 

This task does not contribute to any scores but serves several functions. First, it provides a comfortable introduction to the interactive mode of the written test as a whole. Second, it allows candidates to familiarize themselves with the keyboard. Third, the candidate's response appears on the score report, allowing test score users to view a sample of the candidate's writing.

Part B: Sentence Completion

In this task, candidates read a sentence that has a word missing, and they supply an appropriate word to complete the sentence. Occasionally, two adjacent sentences are presented but still only one word is missing. Candidates are given 25 seconds for each item. During this time, candidates must read and understand the sentence, retrieve a lexical item to complete the sentence, and type the word above the line provided. Sentences range in length from 4 to 30 words. Across all items in this task, candidates are exposed to sentences with words missing from various parts of speech (e.g., noun, verb, adjective, adverb) and from different positions in sentences: sentence-initial, sentence-medial, sentence-final.

Examples:

> 1. I'm sorry but your bill is long past _____.
> 2. He arrives _____ and is often the first one here.
> 3. I asked a coworker to take over my _____ because I wasn't feeling well.

It is sometimes thought that fill-in-the-gap tasks (also called cloze tasks) are more authentic when longer passages or paragraphs are presented to the candidate, as this enables context-inference strategies.

However, research has shown that candidates rarely need to look beyond the immediate sentence in order to infer the correct word to fill the gap (Sigott, 2004). This is the case even when test designers specifically design items to ensure that candidates go beyond sentence-level information (Storey, 1997). Readers commonly rely on sentence-level comprehension strategies partly because the sentence surrounding the gap provides clues about the missing word's part of speech and morphology and partly because sentences are the most common units for transmission of written communication and usually contain sufficient context for meaning.

Above and beyond knowledge of grammar and semantics, the task requires knowledge of word use and collocation as they occur in natural language. For example, in the sentence: "The police set up a road _____ to prevent the robbers from escaping," some grammatical and semantically correct words that might fit include "obstacle", "blockage" or "impediment." However, these would seem inappropriate word choices to a native reader, whose familiarity with word sequences in English would lead them to expect a word such as "block" or "blockade."

In many Sentence Completion items there is more than one possible correct answer choice. However, all items have been piloted with native speakers and learners of English and have been carefully reviewed with reference to content, collocation and syntax. The precise nature of each item and possible answer choices are quantified in the scoring models.

The sentence completion task draws on interpretation, inference, lexical selection and morphological encoding, and as such contributes to candidate's Word Choice (Writing) and Reading scores.

Part C: Passage Reconstruction

Passage Reconstruction is similar to a task known as free-recall, or immediate-recall: Candidates are required to read a text, put it aside, and then write what they can remember from the text. In this task, a short passage is presented for 30 seconds, after which the passage disappears and the candidate has 90 seconds to reconstruct the content of the passage in writing. Passages range in length from 30 to 75 words. The items sample a range of sentence lengths, syntactic variation and complexity. Two discourse genres are presented in this task: narrative and email. Narrative texts are short stories about common situations involving characters, actions, events, reasons, consequences, or results. Email texts are adapted from authentic electronic communication and may be conversational messages to colleagues or more formal messages to customers.

Examples:

(Narrative) Corey is a taxi driver. It is his dream job because he loves driving cars. He started the job ten years ago and has been saving up money since then. Soon, he will use this money to start his own taxi company.

(E-Mail) Thank you so much for being so understanding about our delay of shipment. It has been quite difficult to get materials from our suppliers due to the recent weather conditions. It is an unusual circumstance. In any case, we should be able to ship the products to you tomorrow. In the meantime, if you have any questions, please feel free to contact me.

In order to accurately reconstruct a passage, the candidate must read the passage presented, understand the concepts and details, and hold them in short-term memory in order to reconstruct the passage. Individual candidates may naturally employ different strategies when performing the task. Reconstruction may be somewhat verbatim in some cases, especially for shorter passages answered by advanced candidates. For longer texts, reconstruction may be accomplished by paraphrasing and drawing on the candidate's own choice of words. Regardless of strategy, the end result is evaluated based on the candidate's ability to reproduce the key points and details of the source passage using grammatical and appropriate writing. The task requires the kinds of skills and core language competencies that are necessary for activities such as responding to requests in writing, replying to emails, documenting events or decisions, summarizing documents, or writing the minutes of meetings.

The Passage Reconstruction task is held to be a purer measure of reading comprehension than, for example, multiple choice reading comprehension questions, because test questions do not intervene between the reader and the passage. It is thought that when the passage is reconstructed in the candidate's mother tongue then the main ability assessed is reading comprehension, but when the passage is reconstructed in the target language (in this case, English), it is more an integrated test of both reading and writing (Alderson, 2000:230). Since the Passage Reconstruction task requires appropriate vocabulary usage and accurate production of sentence-level and paragraph-level writing at functional, workplace speeds, the performance of the task is reflected in the Word Choice and Grammar subscores.

Part D: Summary Writing

In the Summary Writing task, candidates read a passage that is between 500 and 750 words in length. The topic of the passage relates to the workplace or to situations that should be familiar to most candidates, including topics such as travel, customer service, human psychology, government and politics, etc. Candidates have 9 minutes to summarize the passage; the written summary must be between 40 and 60 words in length.

Writing an effective summary requires reading and understanding the text, determining what the overall theme of the passage is (including any underlying position taken by the author), and identifying relevant details. A good summary paraphrases and condenses the contents of the passage to make the topic understood by an unfamiliar reader. Therefore, a good summary synthesizes the passage and relates (only) the most important supporting details using correct and comprehensible English. For these reasons, the summary writing task contributes to Reading and Grammar (Writing) scores. Because a good summary does not involve simply copying portions of the passage directly, candidates are penalized if too much of their response contains text lifted verbatim from the passage.

Example:

> Doing business abroad can be complicated. There are many more factors to consider than when doing business in one's home country. Language, culture, and customs vary from country to country. Knowledge of these is crucial to a company's long-term success. Without cultural awareness, communication may be difficult and business opportunities may be lost.
>
> Most people think that language differences are the main cause of cross-cultural miscommunication in the workplace. It is true that speaking different languages can make business dealings more difficult. However, professionals can easily overcome linguistic barriers by hiring a skilled interpreter. Nevertheless, it is impossible for people to communicate effectively if they don't understand each other's cultural background. For example, in many Western countries, people are very direct in their communication with one another. They express their concerns and expectations clearly and explicitly. For business people from some Eastern countries, though, such directness can be viewed as rude or insulting. Westerners traveling to the East can avoid offending their hosts by being aware of such cultural expectations. Having knowledge about a country's cultural expectations can help business people interpret the behavior of their foreign counterparts. It also helps business people understand how their own behavior might be interpreted by others.
>
> Culture determines much more than acceptable behaviors. It also impacts the specific needs of consumers in a particular region. Local customs dictate what type of foods people eat, as well as what products or services they use regularly. Because of this, it is crucial for companies doing business abroad to understand the local customs and tailor their business strategies to the local market. For instance, a fast food company with restaurants in nearly 120 countries took beef off the menu when it opened its doors for business in India, where eating beef is taboo. Half the restaurants' customers in India are also vegetarian, so the company added vegetarian dishes flavored with Indian spices. In addition, the company had to teach customers what it meant to be a self-service restaurant and that they needed to walk up to the counter to order food. The company's success with these strategies shows the value of understanding the link between culture and market demands.
>
> The day-to-day practice of running a business can vary greatly from culture to culture. For example, in some places, workers are expected to work eight hours with only a few short breaks during the day. In contrast, employees from other regions are used to a more relaxed workday with longer breaks. Learning about these expectations is important for business leaders who are planning to open an office or facilities abroad, especially if they are planning to employ foreign workers. Being aware of cultural differences in work practices, and creating a strategy for managing them, can be the key in a company's success. In recent years, many firms have emerged that specialize in helping multinational companies do just this. These firms provide valuable cultural awareness workshops for new employees. According to reports, employee retention is higher among those who participate in such workshops.
>
> People also like to feel that they are getting something for nothing. Consumers prefer knowing that spending now will result in a reward of some kind at a later time. Even when the value of the gift or discount is not very large, consumers respond strongly to incentives. Marketing departments know how strongly people respond to incentives and use them to encourage people to spend their money.

Part E: Reading Comprehension

In this task, candidates are presented with a passage and two questions with multiple choice options. The passage consists of written material drawn from everyday or workplace situations. The passage and options may include graphs or charts, but such figures are very basic and require only a limited amount of graphic literacy to understand.

Example:

(A) Tulip Financial Group offers industry-leading financial services. (B) A new office has just opened in Orlando. (C) We provide customized knowledge to guide our clients. (D) Our outstanding representatives help you determine which opportunities are right for you. Our knowledge base can help you get a big advantage in today's market. Our services include:

- Strategic planning
- Comprehensive reports that allow comparisons across industries and companies
- Customized reports including key issues in organizations of interest to you
- Frequent market summaries and trends
- Up-to-date trading data on publicly traded organizations

1. What service Tulip Financial Group offer? [FACT]

a. Check cashing and deposit
b. Customized portfolio review
c. Market summaries and trends
d. Personal loan consolidation

2. Choose the sentence that does **not** belong in the passage. [ORGANIZATION]

a. A
b. B
c. C
d. D

The "passage" portion of Reading Comprehension items may contain two portions, which could come from different sources. For example, one text might be an invoice for services rendered while a second could be a letter from a customer disputing a charge. Such passages require integration and synthesis across the two types of text in order to answer questions correctly.

The comprehension questions conform to the following "types", based on which reading skills are required to answer:

- **Main idea**: identify the central theme of the passage
- **Organization**: identify portions of the passage which do not conform to appropriate or logical organization
- **Fact**: locate or verify a particular detail which is explicitly expressed in the text
- **Inference**: answer a question which does not have an implicit referent in the passage

The Reading Comprehension task contributes to the Reading score.

Part F: Email Writing

In this task, candidates are given an opportunity to demonstrate their writing ability using email in relatively formal, work-related settings. Candidates are presented with a short description of a situation and must write an email in response to the situation. Possible functions which candidates might encounter include, but are not limited to: giving suggestions, making recommendations, requesting information, negotiating a problem, giving feedback, and reporting an event. Candidates are given nine minutes to read and respond to the situation. Responses of at least 100 words are expected, and those that are less than 30 words or that are off-topic are assigned the lowest possible score.

Each email situation contains several elements:

- the setting or place of work where the correspondence takes place
- the addressee to whom the email is to be written, and the relationship between the candidate and the addressee
- the goal or functional purpose of the email
- three themes (e.g., suggestions, reasons, or recommendations) which the candidate should address in his/her response

Example:

> You work for a restaurant. The restaurant's manager, Ms. Johnson wants to reward her employees for working hard but can't afford to increase salaries at this time. Write an email to her suggesting three other ways she could reward her staff.
>
> Your suggestions must come from the following three themes:
> - free lunch
> - employee discount
> - vacation days
>
> You should include all three themes. Provide supporting ideas for each of your

Candidates are not expected to generate original content for their responses as the themes to address are provided for them. However, candidates are required to construct elaborations, supporting ideas or reasons for each of the themes. In order to fulfill the task, candidates must understand the situation presented, relate it to their existing knowledge, and synthesize and evaluate the information such that an appropriate response can be composed. Candidates must be conscious of the purpose of the email, address each of the themes, and understand the relationship between themselves as the writer and the intended recipient of the email. Candidates must fully understand the prompt in order to construct an informative, organized, succinct response with appropriate tone, word choice, and grammatical accuracy. Therefore, performance on the Email Writing task is reflected in the Grammar, Word Choice, Voice & Tone and Organization subscores.

Part G: Dictation

In the Dictation task, each item consists of one sentence. When candidates hear a sentence, they must type the sentence exactly as they hear it. Candidates have 25 seconds to type each sentence. The sentences are presented in approximate order of increasing difficulty. Sentences range in length from 3 words to 14 words. The items present a range of grammatical and syntactic structures, including imperatives, wh-questions, contractions, plurals, possessives, various tenses, and particles. The audio item prompts are spoken with a natural pace and rhythm by various native speaker voices that are distinct from the examiner voice.

Examples:

1. There's hardly any paper left.
2. Success is impossible without teamwork.
3. Corporations and companies are staying current with the latest technologies.

Dictation requires the candidate to perform time-constrained processing of the meanings of words in sentence context. The task is conceived as a test of expectancy grammar (Oller, 1971). An expectancy grammar is a system that governs the use of a language for someone who has knowledge of that language. Proficient listeners tend to understand and remember the content of a message, and not the exact words used; they retain the message rather than the words that carry the message. Therefore, when writing down what they have heard, candidates need to use their knowledge of the language either to retain the word string in short-term memory or to reconstruct the sentence that they have forgotten. Those with good knowledge of English words, phrase structures, and other common syntactic forms can keep their attention focused on meaning, and fill in the words or morphemes that they did not attend to directly in order to reconstruct the text accurately (Buck, 2001:78).

The task provides information on comprehension, language processing, and writing ability. As the sentences increase in length and complexity, the task becomes increasingly difficult for candidates who are not familiar with English words and sentence structures. Analysis of errors made during dictation reveals that the errors relate not only to interpretation of the acoustic signal and phonemic identification, but also to communicative and productive skills such as syntax and morphology (Oakeshott-Taylor, 1977). For these reasons, the Dictation task contributes to Grammar (Writing) and Listening scores.

Part H: Response Selection

In the Response Selection task, candidates listen to a sentence, which is immediately followed by three possible responses. From among the three possible responses, candidates choose the one that is the most appropriate response to the sentence. Candidates answer each question either by clicking 'A', 'B', or 'C'.

Example:

Our profit last year was higher than expected.

A: Let's celebrate.
B: That's too bad.
C: We lost a lot last year.

The sentences and possible responses are spoken at a conversational pace. This task is designed to measure candidates' listening comprehension ability. The task demands immediate word recognition and extraction of meaning in the stream of speech, comprehension of the key proposition in the sentence and identification of which response is the best match given the sentential context.

Part I: Conversations

In the Conversations task, candidates listen to a conversation between two speakers, which typically consists of three short sentences. Immediately after the conversation, an examiner voice asks a comprehension question and candidates answer the question with a word or short phrase.

Example:

> Speaker 1: How was the business trip?
> Speaker 2: There was a storm the whole time.
> Speaker 1: That sounds terrible.
>
> What happened during the business trip?

This task measures candidates' listening comprehension ability. Conversations are recorded at a conversational pace covering a range of topics. The task requires candidates to follow speaking turns and extract the topic and content from the interaction at a conversational pace. Quick word recognition and decoding and efficient comprehension of meaning are critical in correctly answering the question.

Part J: Passage Comprehension

In the Passage Comprehension task, candidates listen to a spoken passage (usually a story) and then are presented with three comprehension questions about the passage. The passages range from 40 to 70 words in length. Most passages are simple stories with a situation involving a character (or characters), a setting, and an ending. The body of the story typically describes an action performed by the agent of the story followed by a possible reaction or implicit sequence of events. The ending typically introduces a result, new situation, actor, patient, thought, or emotion.

Example:

> Jason woke up feeling sick. He called his boss and explained that he could not come in to work. Immediately after making the phone call, he took some medicine. A few hours later, Jason no longer felt sick. Rather than waste the afternoon at home, he decided to go to work after all.
>
> *After listening to a passage, the candidate hears and responds to three comprehension questions.*
>
> Question 1: What problem did Jason have when he woke up?
> Question 2: What did he do right after calling his boss?

For each passage, candidates are asked to answer three comprehension questions. Correct answers to the questions (or information needed for simple inferences) are all included in the passage. Questions typically ask for the main idea and details of the passage. Unlike Response Selection and Conversation, the Passage Comprehension task allows for the assessment of candidates' listening comprehension ability with longer speech.

Part K: Passage Reading

In the Passage Reading task, candidates are asked to read two short passages out loud, one at a time. Candidates are given 30 seconds to read each passage. The reading texts are printed on the test paper or displayed on the computer screen.

The passages take the form of either an expository text or an email message and deal with typical business topics or activities. All passages are relatively simple in structure and vocabulary and range in length from 40 to 55 words. The SMOG Readability Index (http://www.harrymclaughlin.com/SMOG.htm) was used to identify and refine the readability score for each passage. SMOG estimates the number of years of education needed to comprehend a passage. The algorithm factors in the number of polysyllabic words across sentence samples (McLaughlin, 1969). All passages have a readability score between 9 and 12, which is at a high school level. They can be read easily and fluently by most educated English speakers.

Examples:

1. Many companies are becoming more and more diverse in the current global market. Some companies encourage diversity in their workplace. The key to a successful work environment is to appreciate each other's background. The goal is to embrace diversity rather than deny differences between people.

2. We have several offices for rent in a large office building. The building is surrounded by trees. All offices have private balconies and hardwood floors. There are many features including an outdoor eating area and a shower. The location is within a few steps of many shops and cafes.

For candidates with little facility in spoken English but with some reading skills, this task provides samples of their pronunciation and oral reading fluency. In addition to information on reading rate, rhythm, and pronunciation, the scoring of the Passage Reading task is informed by miscues (Goodman, 1969). Miscues occur when a reading is different from the words on the page or screen, and provide information about how well candidates can make sense of what they read. For example, hesitations or word substitutions are likely when the decoding process falters or cannot keep up with the current reading speed; word omissions are likely when meaning is impaired or interrupted. More experienced readers draw on the syntax and punctuation of the passage, as well as their knowledge of commonly co-occurring word patterns; they can monitor their rate of articulation and comprehension accordingly. This ability to monitor rate helps ensure that reading is steady as well as rhythmic, with correct stress and intonation that conveys the author's intended meaning. Less experienced readers are less able to comprehend, articulate and monitor simultaneously, resulting in miscues and breaks in the flow of reading.

Part L: Repeats

In this task, candidates are asked to repeat sentences verbatim. The administration is interactive. The system plays a sentence spoken by a native speaker and the candidate attempts repeating it; then the system plays another sentence and the candidate repeats it. The interaction continues in this way until the candidate completes the section. The sentences are presented to the candidate in approximate order of increasing difficulty. Sentences range in length from 3 words to 15 words. The audio item prompts are spoken in a conversational manner.

Examples:

1. It took a lot longer than expected.
2. Come to my office after class if you need help.
3. People know how easy it is to get lost in thought.

To repeat a sentence longer than about seven syllables, a person must recognize the words as spoken in a continuous stream of speech (Miller & Isard, 1963). Highly proficient speakers of English can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with English words, phrase structures, and other common syntactic forms. If a person habitually processes five-word phrases as a unit (e.g. "the really big apple tree"), then that person can usually repeat utterances of 15 or 20 words in length. Generally, the ability to repeat material is constrained by the size of the linguistic unit that a person can process in an automatic or nearly automatic fashion. As the sentences increase in length and complexity, the task becomes increasingly difficult for speakers who are not familiar with English sentence structure.

Because the Repeat items require candidates to organize speech into linguistic units, Repeat items assess the candidate's mastery of phrase and sentence structure. Given that the task requires the candidate to repeat full sentences (as opposed to just words and phrases), it also offers a sample of the candidate's fluency and pronunciation in continuous spoken English.

Part M: Sentence Builds

In the Sentence Builds task, candidates hear three short phrases and are asked to rearrange them to make a sentence. The phrases are presented in a random order (excluding the original word order), and the candidate is expected to say a reasonable and grammatical sentence that comprises exactly the three given phrases.

Examples:

1. my boss / to California / moved
2. the prices range / to thirty dollars / from fifteen
3. to their leader / listened carefully / the young men

To correctly complete this task, a candidate must understand the possible meanings of the phrases and know how they might combine with other phrasal material, both with regard to syntax and pragmatics. The length and complexity of the sentence that can be built is constrained by the size of the linguistic unit (e.g., one word versus a three-word phrase) that a person can hold in verbal working memory. This is important to measure because it reflects the candidate's ability to access and retrieve lexical items and to build phrases and clause structures automatically. The more automatic these processes are, the more the candidate's facility in spoken English. This skill is demonstrably distinct from memory span (see Section 3, Test Construct).

The Sentence Builds task involves constructing and articulating entire sentences. As such, it is a measure of candidates' mastery of sentences in addition to their pronunciation and fluency.

Part N: Speaking Situations

In this task, candidates listen to and read a brief scenario and are then asked to respond as if they were in the situation. Candidates have 10 seconds to prepare a response and 40 seconds to respond to each situation. Candidates are expected to give pragmatically appropriate responses as well as respond using accurate grammar and appropriate connectors and cohesive devises.

Example:

> You borrowed a jacket from your friend, Mark. However, you spilled coffee on it, and it left a large stain. Mark calls and says he needs his jacket. What would you say to him?

The Speaking Situations task elicits aspects of pragmatic ability in a relatively open, long turn response. Candidates must demonstrate awareness and appropriate use of the kind of language required in different social situations eliciting speech acts such as apologizing, requesting, refusing, etc. Responses provide information about candidates' pronunciation, fluency and vocabulary. In addition, responses are scored based on the appropriateness and clarity of the response for the given situation, the effectiveness and extent to which the social demand is conveyed, and the extent to which the candidate used appropriate politeness conventions and spoken register.

Part O: Story Retelling

In this task, candidates listen to a brief story and are then asked to describe what happened in their own words. Candidates have 30 seconds to respond to each story. Candidates are encouraged to tell as much of the story as they can, including the situation, characters, actions and ending. The stories consist of three to six sentences and contain from 30 to 90 words. The situation involves a character (or characters), setting, and goal. The body of the story describes an action by the agent of the story followed by a possible reaction or implicit sequence of events. The ending typically introduces a new situation, actor, patient, thought, or emotion.

Example:

> Paul planned on taking the late flight out of the city. He wasn't sure whether it would be possible because it was snowing quite hard. In the end, the flight was cancelled because there was ice on the runway.

The Story Retelling items assess a candidate's ability to listen and understand a passage, reformulate the passage using his or her own vocabulary and sentence structure, and then retell it in detail. This section elicits longer, more open-ended speech samples than earlier sections in the test, and allows for the assessment of a wide range of spoken abilities. Performance on Story Retelling provides a measure of sentence mastery, vocabulary, fluency, and pronunciation.

3. Test Construct

3.1 Facility in English

For any language test, it is essential to define the test construct, or the skills and knowledge reflected in the test scores (Bachman, 1990; Bachman & Palmer, 1996). E^Pro is designed to measure a candidate's facility in English in the workplace context, which is how well the candidate can understand spoken or

written English and respond appropriately in written or spoken English on everyday and workplace topics at a functional pace.

3.1.1 Facility in Written English

The constructs that can be observed in the candidate's performances in E^Pro are knowledge of the language, such as grammar and vocabulary, and knowledge of writing conventions, such as organization and tone. Underlying these observable performances are psycholinguistic skills such as automaticity and anticipation. As candidates operate with texts and select words for constructing sentences, those who are able to draw on many hours of relevant experience with grammatical sequences of appropriate words will perform at the most efficient speeds.

The first concept embodied in the definition of facility is how well a candidate understands spoken or written English. Both modalities of encoding (listening and reading) are covered in the test. The Dictation task exposes candidates to spoken English and the remaining sections present written English that candidates must read and comprehend within given time limits.

Listening dictation requires segmenting the acoustic stream into discrete lexical items and receptively processing spoken language forms including morphology, phrase structure and syntax in real-time. The task simulates use of the same skills that are necessary for many real-life written tasks, such as professional transcribing, listening to a customer over the telephone and inputting information into an electronic form, and general listening and note-taking. Buck (2001) asserts that dictation is not so much an assessment of listening skills, as it is sometimes perceived, but rather an assessment of general language ability, requiring both receptive and productive knowledge. This is because it involves both comprehension and (re)production of accurate language.

Reading requires fluent word recognition and problem-solving comprehension abilities (Carver, 1991). Interestingly, the initial and most simple step in the reading process, word recognition, is what differentiates native readers from even highly proficient second-language readers (Segalowitz et. al., 1991). Native readers have massively over-learned words by encountering them in thousands of contexts, which means that they can access meanings automatically and also anticipate frequently-occurring surrounding words.

Proficient language users consume fewer cognitive resources when processing spoken English or analyzing English text visually, and therefore have capacity available for other higher-level comprehension processes. Comprehension is conceived as parsing sentences, making inferences, resolving ambiguities, and integrating new information with existing knowledge (Gough et. al., 1992). Alderson (2000:43) suggests that these comprehension skills involve vocabulary, discourse and syntactic knowledge, and are therefore general linguistic skills which may pertain to listening and writing as much as they do to reading.

By utilizing integrated listening/reading and written response tasks, E^Pro taps core linguistic skills and measures the ability to understand, transform and rework texts. After initial identification of a word, either as acoustic signal or textual form, candidates who are proficient in the language move on to higher-level prediction and monitoring processes including anticipation. Anticipation enables faster and more accurate decoding of language input, and also underlies a candidate's ability to select appropriate words when producing text. The key skill of anticipation is assessed in the Sentence Completion and Passage Reconstruction tasks of the E^Pro exam as candidates are asked to anticipate missing words and reconstruct textual messages.

The second concept in the definition of facility in written English is how well the candidate can respond appropriately in writing. The composition tasks in E^Pro are designed to assess not only proficiency in the core linguistic skills of grammatical and lexical range and accuracy, as described above, but also the other essential elements of good writing such as organization, effective expression of ideas, and voice. These are not solely language skills but are more associated with effective writing and critical thinking, and must be learned. Assuming these skills have been mastered in the writer's first language (L1), they may be transferable and applied in the writer's L2, if their core linguistic skills in L2 are sufficiently advanced. Skill in organization may be demonstrated by: presenting information in a logical sequence of ideas; highlighting salient points with discourse markers; signposting when introducing new ideas; giving main ideas before supporting them with details. When responding to an email, skill in voice and tone may be demonstrated by: properly addressing the recipient; using conventional expressions of politeness; showing understanding of the recipient's point of view by rearticulating their opinion or request; and fully responding to each of the recipient's concerns.

Because the most widely used form of written communication is email, E^Pro directly assesses the ability to compose informative emails with accuracy and correct word choice, while also adhering to the modern conventions regarding style, rhetoric, and degree of formality for business settings.

The last concept in the definition of facility in written English is the candidate's ability to perform the requested tasks at a functional pace. The rate at which a candidate can process spoken language, read fluently, and appropriately respond in writing plays a critical role in whether or not that individual can successfully communicate in a fast-paced work environment. A strict time limit imposed on each item ensures that proficient language users are advantaged and allows for discriminating candidates with different levels of automaticity.

The scoring of E^Pro is grounded in research in applied linguistics. A taxonomy of the components of language knowledge which are relevant to writing are presented in a model by Grabe and Kaplan (1996). Their model divides language knowledge into three types: linguistic knowledge, discourse knowledge, and sociolinguistic knowledge. These are broadly in line with the E^Pro subscores of Grammar and Word Choice (linguistic knowledge), Organization (discourse knowledge), and Voice & Tone (sociolinguistic knowledge).

Table 3. Taxonomy of Language Knowledge (adapted and simplified from Grabe and Kaplan, 1996: 220-221).

Knowledge	Description
1. Linguistic Knowledge	a. Written code (spelling, punctuation) b. Phonology and morphology (sound/letter correspondence, morpheme structure) c. Vocabulary (interpersonal, academic, formal, technical, topic-specific, non-literal words and phrases) d. Syntactic/Structural (syntactic patterns, formal structures, figures of expression)
2. Discourse Knowledge	a. Marking devices (cohesion, syntactic parallelism) b. Informational structuring (topic/comment, given/new) c. Recognizing main topics d. Organizing schemes (top-level discourse structure) e. Inferencing (bridging, elaborating)
3. Sociolinguistic Knowledge	a. Functional uses of written language b. Register and situation (status of interactants; degree of formality; degree of distance; topic of interaction) c. Sociolinguistic awareness across languages and cultures

Aligned with the taxonomy presented in Table 3, linguistic knowledge maps onto a linguistic aspect of performance in the scoring of the test; whereas discourse and sociolinguistic knowledge relate to a rhetoric aspect. Comprehension is not mapped explicitly onto the taxonomy because it addresses language knowledge as opposed to the specific information conveyed by the language. However, comprehension is recognized as an important factor for facility in written English, and is, therefore, identified as a unique aspect of the candidate's performance in the scoring.

In sum, there are many processing elements required to participate in a written exchange of communication: a person has to recognize spoken words or words written in an email or text received, understand the message, formulate a relevant response, and then compose stylistically appropriate sentences. Accordingly, the constructs that can be observed in the candidate's performances in E^Pro are knowledge of the language, such as grammar and vocabulary, comprehension of the information conveyed through the language, and knowledge of writing conventions, such as organization and tone. Underlying these observable performances are psycholinguistic skills such as *automaticity* and *anticipation*. As candidates operate with texts and select words for constructing sentences, those who are able to draw on many hours of relevant experience with grammatical sequences of appropriate words will perform at the most efficient speeds.

3.1.2 Facility in Spoken English

E^Pro also measures a candidate's *facility in spoken English – that is the ability to understand spoken English on everyday and workplace topics and to respond appropriately at a native-like conversational pace in intelligible English*. Another way to express the construct, *facility in spoken English*, is "ease and immediacy in understanding and producing appropriate conversational English" (Levelt, 1989). There are many processing elements required to participate in a spoken conversation: a person has to track what is being said, extract meaning as speech continues, and then formulate and produce a relevant and intelligible response. These component processes of listening and speaking are schematized in Figure 2, adapted from Levelt (1989).

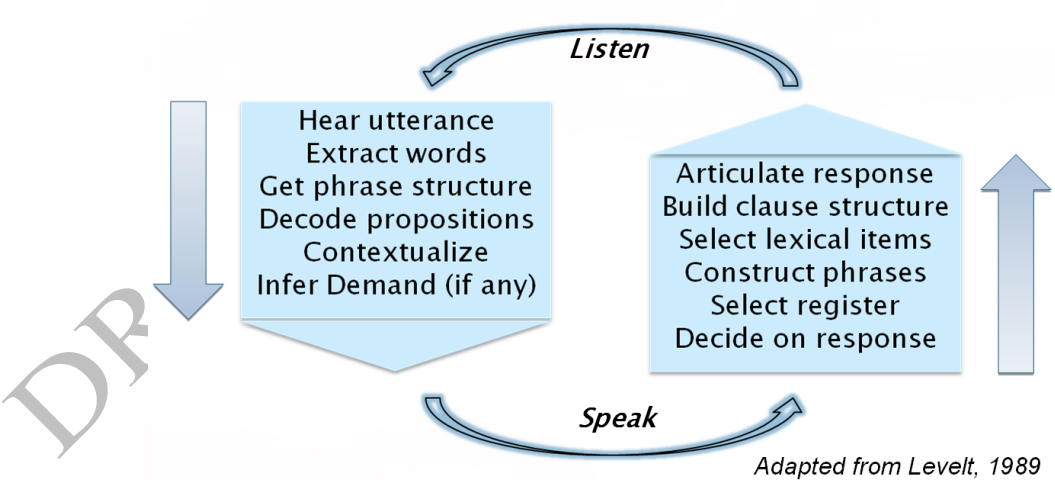

Figure 2. Conversational processing components in listening and speaking.

Core language component processes, such as lexical access and syntactic encoding, typically take place at a very rapid pace. During spoken conversation, Van Turennout, Hagoort, and Brown (1998) found that speakers go from building a clause structure to phonetic encoding in about 40 milliseconds. Similarly, the other stages shown in Figure 1 have to be performed within the small period of time available to a

speaker involved in interactive spoken communication. A typical window in turn taking is about 500-1000 milliseconds (Bull and Aylett, 1998). If language users cannot perform the internal activities presented in Figure 1 in real time, they will not be able to participate as effective listener/speakers. Thus, spoken language facility is essential in successful oral communication.

Automaticity in language processing is the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate responses without conscious attention to the linguistic code (Cutler, 2003; Jescheniak, Hahne, & Schriefers, 2003; Levelt, 2001). Automaticity is required for the speaker/listener to be able to focus on what needs to be said rather than to how the language code is structured or analyzed. By measuring basic encoding and decoding of oral language as performed in integrated tasks in real time, E^Pro probes the degree of automaticity in language performance.

Three basic types of scores are produced from the test: scores relating to the content of what a candidate says, scores relating to the manner of the candidate's speaking, and scores relating to the candidate's listening proficiency. For the speaking part of the scores (i.e., content and manner), this distinction corresponds roughly to Carroll's (1961) description of a knowledge aspect and a control aspect of language performance. In later publications, Carroll (1986) identified the control aspect as automatization, which occurs when speakers can talk fluently without realizing they are using their knowledge about a language.

The E^Pro exam provides a measurement of the real-time decoding and encoding of spoken English. Performance on E^Pro items predicts a more general spoken English facility, which is essential for successful oral communication in English. The same facility in spoken English that enables a person to satisfactorily understand and respond to the listening/speaking tasks in E^Pro also enables that person to participate in native-paced conversation.

3.2 The Role of Context

Grabe and Kaplan's taxonomy explains why some of the test material is context-independent (e.g. Sentence Completion) and some material is context-bound. Scoring related to Linguistic Knowledge, such as vocabulary, discourse and syntactic knowledge, can be elicited from performance on context-bound material but is more efficiently elicited from performance on context-independent material. Scoring related to Discourse and Sociolinguistic Knowledge, however, requires context, awareness of audience, and functional purpose for communication.

In general, E^Pro items present context-independent material in English. Context-independent material is used in the test items for three reasons. First, context-independent items exercise and measure the most basic meanings of words, phrases, and clauses on which context-dependent meanings are based (Perry, 2001). Second, when language usage is relatively context-independent, task performance depends less on factors such as world knowledge and cognitive style and more on the candidate's facility with the language itself. Thus, the test performance relates most closely to language abilities and is not confounded with other candidate characteristics. Third, context-independent tasks maximize response density; that is, within the time allotted for the test, the candidate has more time to demonstrate performance in writing the language because less time is spent presenting contexts that situate a language sample or set up a task demand.

The two exceptions to this context-independence are the following tasks: Speaking Situations and Email Writing. The Email Writing task presents a situation with schema that candidates must attune to, for example, the purpose of the writing and the relationship between themselves and the intended recipient of the email. In this way, Email Writing allows for the assessment of the grammar and mechanics of

writing, as well as knowledge of the email genre and the rhetorical and cultural norms for organizing information in emails. The Speaking Situations task similarly presents a situation requiring candidates to infer the sociolinguistic demand and produce an appropriate response that would successfully satisfy the demand. In both cases, candidates are provided with contextual information to allow them to build schema. Similarly, for both of these tasks, achieving a high score requires the ability to employ sociolinguistic knowledge (i.e., in this situation, what kinds of language, linguistic structures, and tone are appropriate?) and to convey the appropriate sociocultural message using spoken (Speaking Situations) or written (Email Writing) English.

All spoken item content is designed to be region-neutral. The content specification also requires that both native speakers and proficient non-native speakers find the items easy to understand and to respond to appropriately. For English learners, the items probe a broad range of skill levels and skill profiles.

For most item types, the E^Pro exam probes the psycholinguistic elements of language performance rather than the social, rhetorical, and cognitive elements of communication. The reason for this focus is to ensure that test performance relates most closely to the candidate's facility with the language itself and is not confounded with other factors. The goal is to separate familiarity with language from other types of knowledge including cultural familiarity, understanding of social relations and behavior, and the candidate's own cognitive style. Also, by focusing on context-independent material, less time is spent developing a background cognitive schema for the tasks, and more time is spent collecting data for language assessment.

3.3 The Role of Memory

Some measures of language proficiency can be misconstrued as memory tests. Since some E^Pro tasks involve holding sentences or situations in memory in order to type or repeat them, or re-assembling paragraphs into reasonable sentences from memory, it may seem that these tasks are unduly influenced by general memory performance. This concern is mitigated by the fact that all relevant items have been presented to samples of educated native speakers of English, and at least 85% of the speakers in that educated native speaker sample responded correctly. If memory, as such, were an overriding component of performance on E^Pro tasks, then native English speakers should show greater performance variation on these items according to the presumed range of individuals' memory spans (see Section 8.2.6 for native-speaker performance). Also, if memory capacity (rather than language ability) were a principal component of the variation among people performing these tasks, the test would not correlate so closely with other accepted measures of language proficiency (see Section 8.3 for TOEIC and 8.4 for CEFR Level Estimates).

4. Content Design and Development

4.1 Vocabulary Selection

The vocabulary used in the test items was taken from a general English corpus and a business English word list. The general English corpus was restricted to forms of the 8,000 most frequent words found in the Switchboard Corpus (Godfrey and Holliman, 1997), a corpus of three million words taken from spontaneous telephone conversations. The business English word list was restricted to forms of the 3,500 most frequent words found in the *University of Cambridge Business English Certificate Preliminary Wordlist*, *Barron's 600 Essential Words for the TOEIC*, and *Oxford Business and Finance words*.

4.2 Item Development

E^Pro items were drafted by trained item writers. All items writers have advanced degrees or training in applied linguistics, TESOL, or language testing. In general, structures used in the test reflect those that are used in common everyday or workplace settings. The items employ a wide range of topics from relatively general English domains to common workplace domains. The item writers were provided a list of potential topics/activities/situations with regard to the business domain, such as:

- Announcements
- Business trips
- Complaints
- Customer service
- Fax/Telephone/E-Mail
- Inventory
- Scheduling
- Marketing/Sales

Item writers were specifically requested to write items so that items would not favor candidates with work experience or require any work experience to answer correctly. The items are intended to be within the realm of familiarity of both a typical, educated, native English speaker and an educated adult who has never lived in an English-speaking country.

Draft items were then reviewed internally by a team of test developers, all with advanced degrees in language-related fields, to ensure that they conformed to item specifications and English usage in different English-speaking regions and contained appropriate content. Then, draft items were sent to external experts on three continents. The pool of expert reviewers included several individuals with PhDs in applied linguists and subject matter experts who worked as training and recruitment managers for large corporations. Expert review was conducted to ensure 1) compliance with the vocabulary specification, and 2) conformity with current colloquial English usage in different countries. Reviewers checked that items would be appropriate for candidates trained to standards other than American English.

All items, including anticipated responses for Sentence Completion, were checked for compliance with the vocabulary specification. Most vocabulary items that were not present in the lexicon were changed to other lexical items that were in the corpus and word list. Some off-list words were kept and added to a supplementary vocabulary list, as deemed necessary and appropriate. The changes proposed by the different reviewers were then reconciled and the original items were edited accordingly.

For an item to be retained in the test, it had to be understood and responded to appropriately by at least 85% of a reference sample of educated native speakers of English.

4.3 Item Prompt Recording

4.3.1 Voice Distribution

Thirty native speakers (14 women and 16 men) representing various speaking styles and regions, including the U.S., U.K., and Australia, were selected for recording the spoken prompt materials.

Several non-native speakers also recorded some items. Care was taken to ensure that the non-native speakers were at advanced levels in terms of their speaking ability and that their pronunciation was clear and intelligible. The speakers' country of origin included India, Hong Kong, Taiwan, Korea, and the Netherlands.

Recordings were made in a professional recording studio in Menlo Park, California. In addition to the item prompt recordings, all the test instructions and listening comprehension questions were also recorded by professional voice talents whose voices were distinct from the item voices.

4.3.2 Recording Review

Multiple independent reviews were performed by test developers on all the recordings for quality, clarity, and conformity to natural conversational styles. Any recording in which reviewers noted some type of error was either re-recorded or excluded from installation in the operational test.

5. Score Reporting

5.1 Scoring and Weighting

Of the 107 items in an administration of the English for Professionals Exam, at least 82 responses are used in the automatic scoring. The first items in many sections are considered practice items and are not incorporated into the final score.

The E^Pro score report is comprised of an Overall score, four skill scores (Speaking, Listening, Reading, and Writing), two skill profile scores (Speaking and Writing profiles), and eight analytic subscores, with four coming from Writing (Grammar, Word Choice, Organization, and Voice & Tone) and four coming from Speaking (Sentence Mastery, Vocabulary, Pronunciation, and Fluency). All scores are reported in the range from 100 to 500.

The E^Pro Overall represents the ability to understand English input and provide accurate, appropriate responses at a functional pace for everyday and workplace purposes. It is based on a weighted combination of all four skill scores. Table 4 shows how the four skill scores are weighted to achieve an Overall score.

Table 4. Subscore Weighting in Relation to the E^Pro Overall Score.

Score	Contribution
Speaking	25%
Listening	25%
Reading	25%
Writing	25%
E^Pro Overall Score	100%

Figure 3 illustrates which sections of the test contribute to each of the subscores.

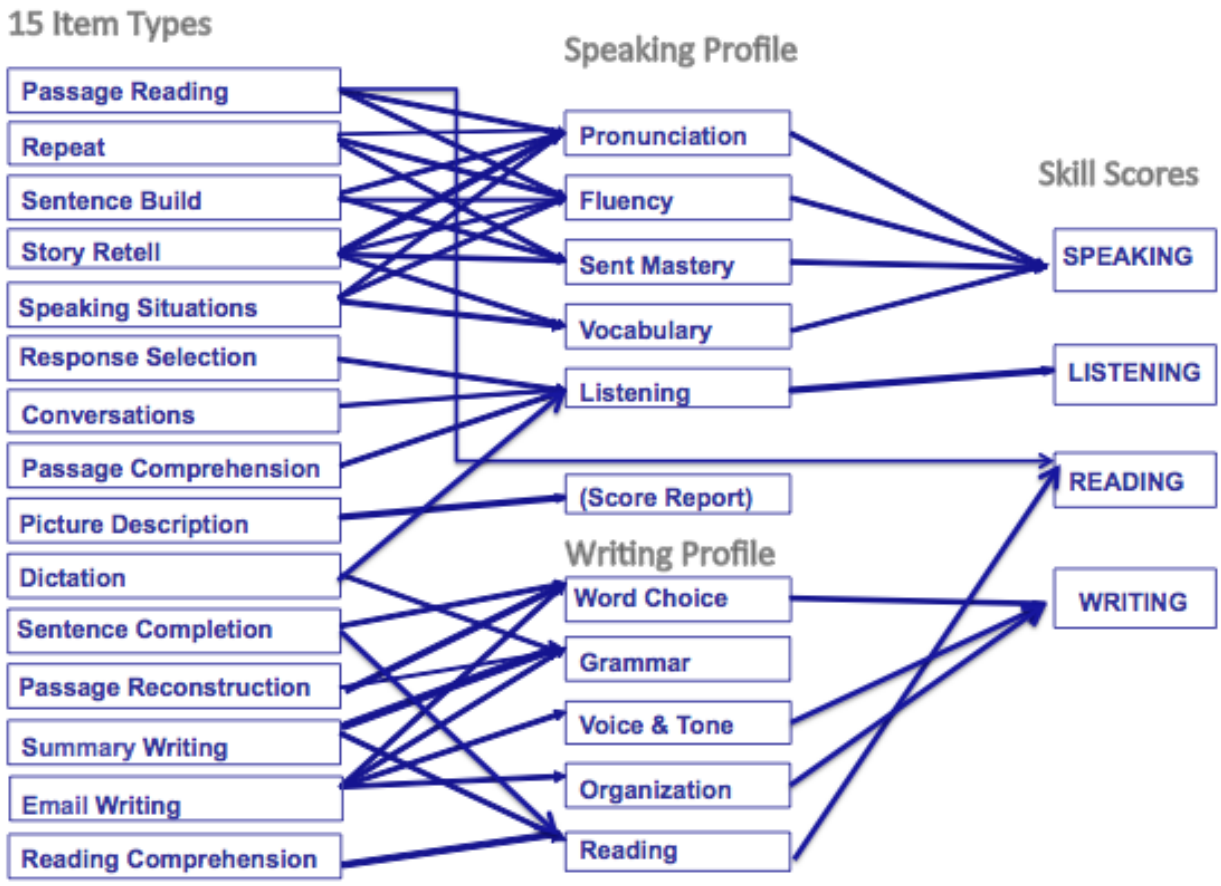

Figure 3. Relation of subscores to item types.

A multi-method, multi-trait approach is taken to ensure that each subscore is reliable and generalizable. The Fluency subscore, for example, is derived from performance on five different tasks. The following sections illustrate how each subscore contributes to Speaking and Writing Profile scores.

5.1.1 Speaking Profile

The sections of the test requiring speaking and listening responses contribute to the subscores which make up the Speaking Profile score.

Sentence Mastery: Sentence Mastery reflects how well the candidate understands and produces a variety of sentence structures in spoken English. The score is based on the ability to use accurate and appropriate words and phrases in meaningful sentences. Sentence Mastery contributes 30% of the Speaking skill score and 20% of the Speaking Profile score.

Vocabulary: Vocabulary reflects how well the candidate understands and produces a wide range of words in spoken English from everyday and workplace situations. The score is based on the familiarity with the meanings of common words and their use in connected

speech. Vocabulary contributes 20% of the Speaking skill score and 20% of the Speaking Profile score.

Fluency: Fluency reflects how well the candidate uses appropriate rhythm, phrasing, and timing when speaking English. The score is based on the ability to speak smoothly and naturally at a conversational pace. Fluency contributes 20% of the Speaking skill score and 20% of the Speaking Profile score.

Pronunciation: Pronunciation reflects how well the candidate produces English consonants, vowels, words and phrases in an intelligible, native-like manner. The score is based on the ability to correctly articulate the sounds of English in connected speech. Pronunciation contributes 30% of the Speaking skill score and 20% of the Speaking Profile score.

Listening: Listening reflects how well the candidate understands specific details and main ideas from everyday and workplace English speech. The score is based on the ability to track meaning and infer the message from English that is spoken at a conversational pace. In addition to contributing 20% of the Speaking Profile score, the Listening score is reported as its own skill score.

Table 5 shows how the five subscores are weighted to achieve the Speaking Profile score on the basis of which the overall Speaking Profile performance description is determined.

Table 5. Subscore Weighting in Relation to Speaking Profile Score.

Score	Contribution
Sentence Mastery	20%
Vocabulary	20%
Fluency	20%
Pronunciation	20%
Listening	20%
Speaking Profile Score	100%

The subscores are based on three different aspects of language performance: a knowledge aspect (the content of a response), a control aspect (the manner in which a response is said), and a comprehension aspect (the extent to which a response reflects the understanding of a listening stimulus). The five subscores reflect these aspects of language performance where Sentence Mastery and Vocabulary are associated with content, Fluency and Pronunciation are associated with manner of speaking, and Listening is associated with comprehension. The content accuracy dimension accounts for 40% of the Speaking Profile score and indicates whether or not the candidate understood the prompt and responded in grammatically accurate sentences and/or with appropriate content. The manner-of-speaking scores count for an additional 40% of the Speaking Profile score, and indicate whether or not the candidate speaks in a native-like manner. The remaining 20% of the Speaking Profile score comes from the comprehension score and indicates whether or not the candidate understood the spoken material. In smooth, successful communication, it is essential to be able to understand what is being said or asked in a stream of speech. Furthermore, producing accurate lexical and structural content is important, but excessive attention to accuracy can lead to disfluent speech production and can also hinder oral communication; on the other hand, inappropriate word usage and misapplied syntactic

structures can also hinder communication. Because successful communication depends on these three dimensions, E^Pro is designed to assess each of them.

The Versant automated scoring system scores both the content (including the content of the responses to the listening items) and manner-of-speaking subscores using a speech recognition system that is optimized based on non-native English spoken response data collected during the field test. The content subscores are derived from the correctness of the candidate's response and the presence or absence of expected words in correct sequences. The manner-of-speaking subscores (Fluency and Pronunciation, as the control dimension) are calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. In order to produce valid scores, during the test development stage, these measures were automatically generated on a sample set of utterances (from both native and non-native speakers) and were then scaled to match human ratings.

5.1.2 Writing Profile

The sections of the test requiring reading and producing written responses contribute to the subscores which make up the Writing score.

Grammar: Grammar reflects how well the candidate understands, anticipates and produces a variety of sentence structures in written English. The score is based on the ability to use accurate and appropriate words and phrases in meaningful sentences. Grammar contributes 40% of the Writing skill score and 25% of the Writing Profile score.

Word Choice: Word choice reflects how well the candidate understands and produces a wide range of words in written English from everyday and workplace situations. The score is based on accuracy and appropriateness of word use for topic, purpose, and audience. Word Choice contributes 30% of the Writing skill score and 25% of the Writing Profile score.

Organization: Organization reflects how well the candidate presents ideas and information in written English in a clear and logical sequence. The score is based on the ability to guide readers through written text and highlight significant points using discourse markers. Organization contributes 15% of the Writing skill score and 10% of the Writing Profile score.

Voice & Tone: Voice and Tone reflects how well the candidate establishes an appropriate relationship with the reader by adopting an appropriate style and level of formality. The score is based on the writer's ability to address the reader's concern and have an overall positive effect. Voice and Tone contributes 15% of the Writing skill score and 10% of the Writing Profile score.

Reading: Reading reflects reflects how well the candidate understands written English texts on everyday and workplace topics. The score is based on the ability to operate at functional speeds to extract meaning, infer the message, and respond appropriately. In addition to contributing 30% of the Writing Profile score, the Reading score is reported as its own skill score.

Table 6 shows how the subscores are weighted to achieve a Writing Profile score on the basis of which the overall Writing Profile performance description is determined.

The subscores are based on several aspects of the candidate's performance: a linguistic aspect (the range and accuracy of word use), a content aspect (the comprehensiveness of the information given), and a rhetoric aspect (the organization and presentation of information).

Table 6. Subscore weighting in relation to Writing Profile score.

Score	Contribution
Grammar	25%
Word Choice	25%
Organization	10%
Voice & Tone	10%
Reading	30%
Writing Profile Score	100%

The linguistic aspect is informed by the Grammar and Word Choice subscores. Combined, these two dimensions account for 50% of the overall score because knowledge of a wide range of words and the accuracy of their use are the pre-requisites of successful written communication. If a candidate is unable to produce coherent sentences that convey the intended meaning in English, then the other dimensions of content and rhetoric may be of limited value. Conversely, if a candidate is strong in the mechanical skills of written language, then s/he has a foundation upon which to learn higher order comprehension and rhetorical skills.

The content aspect, or comprehensiveness of the information given in a candidate's response, is associated with the Reading subscore. This accounts for 30% of the Writing Profile score. It is not only a measure of how well the candidate is able to understand textual input, but also how well the candidate then demonstrates understanding by responding to it. Thus, this is not a measure of pure comprehension in the cognitive sense, but rather of comprehension and usage.

Finally, the rhetoric aspect is informed by the Organization and Voice & Tone subscores. This aspect also accounts for 20% of the Writing Profile score. Producing accurate lexical and structural content is important, but effective communication depends on producing clear, succinct writing which allows for ease of reading and gives a positive impression to the reader.

5.2 Score Use

Once a candidate has completed a test, the Versant testing system analyzes the performances and makes the scores available through Pearson VUE's PCM portal. Test administrators and score users can then view and print out the test results from a password-protected section of the website.

Score users of E^Pro may be business organizations, educational and government institutions. Business organizations may use E^Pro scores as part of the screening, hiring, selection, language monitoring or promotion process. Within a pedagogical research setting, E^Pro scores may be used to evaluate the level of English proficiency of individuals entering into, progressing through, and leaving English language courses.

The E^Pro score scale covers a wide range of abilities in spoken and written English communication. In most cases, score users must decide what score is considered a minimum requirement in their context (i.e., a cut score). Score users may wish to base their selection of an appropriate cut score on their own localized research. Pearson can provide a Benchmarking Kit and further assistance in establishing cut scores.

Section II – Field Test and Validation Studies

6. Field Test

6.1 Data Collection

Both native speakers of English and English language learners were recruited as participants from August 2009 through November 2009 to take a preliminary data-collection version of the E^Pro exam. Candidates in the data collection process took one or more of three forms of preliminary E^Pro: (1) a modified form which only had Speaking and Listening items (i.e., E^Pro Speaking Profile); (2) a modified form which only had Writing and Reading items (i.e., E^Pro Writing Profile); or (3) a combination of both modified forms.

The purposes of this field testing were 1) to validate operation of the test items with both native speakers and learners, 2) to calibrate the difficulty of each item based on a large sample of candidates at various levels and from various first language backgrounds, and 3) to collect sufficient written and spoken English samples to develop automatic scoring models for the test. The description of participants is presented in Table 7.

Table 7. Description of participants in the field testing whose responses were used to develop automated scoring models for items in Preliminary E^Pro Writing Profile (n=1,695) and Speaking Profile (n=973).

		English Learners	
	Native	**Writing Profile**	**Speaking Profile**
Number of Participants	73	1695	973
Male: Female	31% : 63% Unknown = 6%	44% : 49% Unknown = 7%	50% : 47% Unknown = 3%
Age Range	20 - 73 mean = 35.6	19 - 67 mean = 28.0	19 - 67 mean = 28.9
Languages	English (U.S., U.K., and Australia)	Angami, Arabic, Armenian, Assamese, Bengali, Bhojpuri, Cantonese, Catalan, Cebuano, Chinese, Czech, Dutch, Farsi, Filipino, Fookien, French, Garhwali, German, Gujarati, Haryanvi, Hindi, Italian, Japanese, Kalenjin, Kannada, Korean, Kumani, Lotha, Marathi, Maithali, Malayalam, Manipuri, Mao, Marathi, Nepali, Oriya, Portuguese, Punjabi, Rajasthani, Rongmei, Russian, Serbian, Spanish, Swedish, Tagalog, Taiwanese, Tamil, Telugu, Thai, Turkish, Urdu, Vietnamese, Visayan, Waray-waray, Yoruba	Angami, Arabic, Assamese, Bengali, Cantonese, Catalan, Cebuano, Chavacano, Chinese, Czech, Dutch, Farsi, Filipino, Fookien, French, German, Gujarati, Haryanvi, Hindi, Indonesian, Japanese, Kalenjin, Kannada, Korean, Maithali, Malayalam, Manipuri, Marathi, Marwadi, Oriya, Portuguese, Punjabi, Rongmei, Russian, Spanish, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Urdu, Vietnamese, Visayan, Waray-waray, Yoruba

6.1.1 Native Speakers

A total of 73 educated adult native English speakers were recruited. Most were from the U.S. with a few from the U.K. and Australia. Most of them took the test multiple times producing a total of 706 completed tests. Each test was comprised of a unique set of items, so items did not overlap between the tests. The mean age of the native speaker sample was 35.6 and the male:female ratio was 31:63.

While E^Pro is specifically designed for English learners, responses from native speakers were used to validate the appropriateness of the test items and their performance was also used to evaluate the scoring models.

6.1.2 English Learners

For the Writing Profile version of the preliminary E^Pro, a total of 1695 English language learner candidates were recruited from various countries representing both university students and working professionals. Many of the candidates took both the Speaking Profile version and the Writing Profile version of the preliminary E^Pro, so they may be counted in both columns "Speaking Profile" and "Writing Profile" in Table 7.

A total of 46 countries were represented in the field test, but the majority of the data were collected in Argentina, China, Germany, India, Italy, Japan, Korean, Philippines, Spain, and Taiwan. A total of 55 different languages were reported. The male:female ratio was 44:49 with 7% of the candidates being unreported. The mean candidate age was 28.

For the Speaking Profile version of the preliminary E^pro, a total of 973 non-native candidates were recruited from various countries representing both university students and working professionals. All the data collection tests were taken on the computer platform. Most candidates took both E^Pro Speaking Profile and Writing Profile tests. Almost all the candidates took the test only once.

Based on the country of origin, there were a total of 46 countries represented in the field test, but same as the E^Pro writing profile, the majority of the data were collected in Argentina, China, Germany, India, Italy, Japan, Korean, the Philippines, Spain, and Taiwan. A total of 46 different languages were reported. The male:female ratio was 50:47 with 3% of the candidates being unreported. The mean age was 28.9.

7. Data Resources for Scoring Development

7.1 Data Preparation

During the field test of the preliminary versions of E^Pro, more than 200,000 responses were collected from native speakers and English learners. Subsets of the response data were presented to trained transcribers and raters for developing the automatic scoring models.

7.2 Transcription

Both native and learner responses were transcribed by native speakers of English in order to train an automatic speech recognition system optimized for non-native speech patterns. The majority of the transcribers had a degree in a language-related field such as linguistics, language studies, or English. All transcription work was performed using Pearson's web-based transcription system and following Pearson's transcription annotation guidelines. Prior to the task, transcribers underwent rigorous training on how to use the web-based transcription system and the guidelines. The quality of their transcriptions was closely reviewed for accuracy during the project. A total of 30,718 transcriptions were produced for native responses and 87,746 transcriptions were produced for learner responses.

7.3 Expert Human Rating

Selected item responses to Passage Reconstruction and Email Writing from a subset of candidates were presented to twenty-one educated native English speakers to be judged for content accuracy and vocabulary usage. Selected item responses to Story Retellings from a subset of candidates were presented to nine educated native English speakers to be judged for content accuracy and vocabulary usage to make the Story Retelling task automatically scoreable. Before the raters began rating responses, they were trained to evaluate responses according to analytical and holistic rating criteria. All raters held a master's degree in either linguistics or TESOL.

The raters logged into a web-based rating system and evaluated the written responses to Passage Reconstruction and Email Writing items for such traits as vocabulary, grammar, organization, and voice & tone. They also evaluated transcriptions of Story Retelling responses, one at a time, for content and vocabulary. The raters' judgments were based on transcriptions instead of recorded spoken responses in order to minimize confounding effects - that is, to ensure that pronunciation or fluency qualities would not affect the evaluation of content and vocabulary. Rating stopped when each item had been judged by three raters. For pronunciation and fluency scoring, the models developed for the Versant English Test[1], were used because those pronunciation and fluency models were trained on a much larger sample of English learners and have proven to be very robust and content independent. Both tests are designed to measure facility in spoken English. Empirical evidence has demonstrated that the Versant English Test is a valid tool to assess spoken English.

8. Validation

8.1 Validation Study Design

In this validation section, validity analyses focused on the scores produced for the E^Pro Speaking Profile and Writing Profile of the preliminary E^Pro because they are the foundation of the E^Pro Overall score. The following types of analysis were performed and reported separately in the subsequent sections:

Structural Validity

1. Reliability: whether or not the exam is structurally reliable and assigns scores consistently,
2. Dimensionality: whether or not the different subscores are sufficiently distinct, particularly those within the same "profile" area (e.g., Speaking Profile or Writing Profile),
3. Accuracy: whether or not the automatically scored Preliminary E^Pro scores are comparable to the scores that human listeners and raters would assign,
4. Differentiation among known populations: whether or not preliminary E^Pro scores reflect expected differences and similarities among known populations (e.g., natives vs. English learners),

Concurrent Validity

5. Relation to scores of tests or frameworks with related constructs: how closely do preliminary E^Pro scores predict the reliable information in scores of a well-established English test for a workplace context (i.e., TOEIC); and how do E^Pro scores correspond to the six levels of the Common European Framework of Reference (CEFR)?

[1] Versant English Test is a spoken English test developed by Pearson with abundant empirical evidence demonstrating its validity and reliability.

There are several differences in test structure between the preliminary version of E^Pro used in the studies described below and the current, production version of E^Pro. For example, scores from the preliminary version of E^Pro were reported on a range from 20-80, whereas scores for the production version of E^Pro were slightly revised and transformed to fall onto a broader scale from 100-500. A few tasks which were included in the preliminary version are no longer present in the current version; also, the current version contains three entirely new tasks which were not present in the preliminary version: Speaking Situations, Summary Writing, and Reading Comprehension. To preserve the integrity of the original analyses, and in recognition of the fact that the current scoring of E^Pro includes tasks and scoring models which were absent from the structure of the preliminary version, all scores and validity results reported here are based on the preliminary version of the test, using the 20-80 score scale. However, the psychometric properties of the E^Pro exam are expected to be consistent with, or an improvement upon, those of the preliminary version.

8.1.1 Validation Sample

A total of 124 participants were recruited for a series of validation analyses. These validation participants were recruited separately from the field test candidates. Care was taken to ensure that the training dataset and validation dataset did not overlap for independent validation analyses. This means that the performance samples provided by the validation candidates were excluded from the datasets used for training the scoring models.

Validation subjects were recruited from a variety of countries, first language backgrounds, and proficiency levels and were representative of the candidate population using the preliminary E^Pro. A total of five native speakers were included in the validation dataset. Table 8 summarizes the demographic information of the validation participants.

Table 8. Description of Participants Used to Validate the Scoring Models and Estimate Test Reliability (n=124).

Number of Participants	124 (including 5 native speakers)
Male : Female ratio	44% : 56%
Age Range	19 – 66 mean = 30.4
Languages	Arabic, Chinese, English, Filipino, French, German, Hindi, Italian, Japanese, Korean, Malayalam, Russian, Spanish, Tagalog, Tamil, Telugu, Visayan

8.2 Structural Validity

As mentioned above, the E^Pro Speaking Profile and Writing Profiles scores were separately analyzed to conduct the validation analysis. To understand the consistency and accuracy of E^Pro Speaking Profile and Writing Profile scores and the distinctness of the subscores, the following was examined: descriptive statistics of the validation sample, the standard error of measurement of the preliminary E^Pro Speaking Profile and Writing Profile scores; the reliability of the preliminary E^Pro Speaking Profile and Writing Profile scores (split-half reliability); the correlations between preliminary E^Pro Speaking Profile and Writing Profile scores and its subscores, and between pairs of subscores; comparison of machine-

generated scores of the preliminary E^Pro Speaking Profile and Writing Profile scores with listener-judged scores of the same tests. These qualities of consistency and accuracy of the test scores are the foundation of any valid test.

8.2.1 Descriptive Statistics

The mean Overall score of the validation sample was 51.74 with a standard deviation of 15.27 (on a scale of 20-80) for Writing Profile, and was 49.57 with a standard deviation of 15.15 for Speaking Profile. Table 9 summarizes some descriptive statistics for the validation sample.

Table 9. Descriptive Statistics for the Validation Dataset (n=124).

Measure	Writing Profile	Speaking Profile
Mean	51.74	49.57
Standard Error	1.37	1.36
Median	51.55	48.07
Standard Deviation	15.27	15.15
Sample Variance	233.07	229.54
Kurtosis	-0.44	-0.63
Skewness	0.06	0.33

8.2.2 Standard Error of Measurement

The Standard Error of Measurement (SEM) provides an estimate of the amount of error, due to unreliability, in an individual's observed test score and "shows how far it is worth taking the reported score at face value" (Luoma, 2003: 183). The SEM of the Writing Profile score is 2.2, and the SEM of the Speaking Profile score is 2.3.

8.2.3 Test Reliability

Score reliabilities were estimated by the split-half method. Split-half reliability was estimated for the Overall Writing Profile and Speaking Profile scores and all of the analytic subscores. The split-half method divides a test into two halves and the scores from these two halves are correlated. Then, the correlation coefficient is corrected for full-test reliability using the Spearman-Brown Prophecy Formula. The split-half reliabilities were calculated for both the listener-judged scores and the machine-generated scores. The reliability coefficients are summarized in Table 10 and Table 11.

Table 10 summarizes the split-half reliability results for the Writing Profile scores. It compares the same individual performances, scored by careful human rating in one case and by independent automatic machine scoring in the other case. The values in Table 10 suggest that there is sufficient information in the preliminary E^Pro (Writing Profile) item response set to extract reliable information, and that the effect on reliability of using the Versant automated system, as opposed to careful human rating, is quite small. The high reliability is a good indication that the computerized assessment will be consistent for the same candidate assuming there are no changes in the candidate's language proficiency level.

Table 10. Split-half Reliabilities of Writing Profile scores: Human Scoring versus Machine Scoring (n=124).

Score	Split-half Reliability for Human Scores	Split-half Reliability for Machine Scores
Writing Profile	**0.93**	**0.98**
Grammar	0.97	0.98
Word Choice	0.89	0.91
Organization	0.77	0.87
Voice & Tone	0.79	0.90
Reading	0.92	0.93

The reliability for the Organization and Voice & Tone subscores is lower than the reliability of the other subscores because these subscores are estimated solely from Email Writing, of which only two items are presented in the test. However, the agreement between two raters for these subscores was sufficiently high: inter-rater reliability for Organization was 0.90 and inter-rater reliability for Voice & Tone was 0.93 at the item level (corrected for under-estimation).

Similar to Writing Profile scores, the values in Table 11 suggest that there is sufficient information in the preliminary E^Pro (Speaking Profile) item response set to extract reliable information, and that the effect on reliability of using the Versant speech recognition technology, as opposed to careful human rating, is quite small.

Table 11. Split-half Reliabilities of Speaking Profile scores: Human Scoring versus Machine Scoring (n=124).

Score	Split-half Reliability for Human Scores	Split-half Reliability for Machine Scores
Speaking Profile	**0.99**	**0.98**
Sentence Mastery	0.96	0.92
Vocabulary	0.94	0.88
Fluency	0.99	0.96
Pronunciation	0.99	0.97
Listening	0.93	0.90

8.2.4 Dimensionality: Correlations among Subscores

Ideally, each subscore on a test provides unique information about a specific dimension of the candidate's ability. For language tests, the expectation is that there will be a certain level of covariance between subscores given the nature of language learning. This is due to the fact that when language

learning takes place, the candidate's skills tend to improve across multiple dimensions. However, if all the subscores were to correlate perfectly with one another, then the subscores might not be measuring different aspects of facility with the language.

Table 12 presents the correlations among the Writing Profile subscores and the Writing Profile score for the same validation sample of 124 candidates, which includes five native English speakers.

Table 12. Inter-correlation between Writing Profile Subscores (n=124).

Score	Grammar	Word Choice	Organization	Voice & Tone	Reading	Writing Profile
Grammar	-					0.96
Word Choice	0.81	-				0.96
Organization	0.77	0.81	-			0.89
Voice & Tone	0.79	0.83	0.98	-		0.91
Reading	0.91	0.88	0.87	0.89	-	0.96

As expected, test subscores correlate with each other to some extent by virtue of presumed general covariance within the candidate population between different component elements of written language skills. The Organization and Voice & Tone subscores correlate highly with one another since they are both representing the rhetoric aspect of written language from the same set of items. However, the correlations between the remaining subscores are below unity (i.e., below 1.0), which indicates that the different scores measure different aspects of the test construct.

Table 13 presents the correlations among the Speaking Profile subscores and the Speaking Profile score for the same validation sample of 124 candidates, which includes five native English speakers.

Table 13. Inter-correlation between Speaking Profile Subscores (n=124).

Score	Sentence Mastery	Vocabulary	Fluency	Pronunciation	Listening	Speaking Profile
Sentence Mastery	-					0.91
Vocabulary	0.86	-				0.92
Fluency	0.80	0.80	-			0.94
Pronunciation	0.74	0.82	0.80	-		0.88
Listening	0.77	0.82	0.80	0.71	-	0.91

Again, as expected, test scores correlate with each other to some extent by virtue of presumed general covariance within the candidate population between different component elements of spoken language skills. However, the correlations between the subscores are significantly below unity, which indicates that the different scores measure different aspects of the test construct.

Figure 4 illustrates the relationship between two relatively independent machine scores from the Speaking Profile subscores (Sentence Mastery and Fluency) for the validation sample (n=124). These machine scores are calculated from a subset of responses that are mostly overlapping (Repeats, Sentence Builds, and Story Retellings for Sentence Mastery and Passage Readings, Repeats, Sentence

Builds, and Story Retellings for Fluency). Although these measures are derived from overlapping sets of responses, the subscores clearly extract distinct measures from these responses. For example, candidates with Fluency scores in the 30-50 range have Sentence Mastery scores that are spread roughly evenly over the whole 20-80 score range.

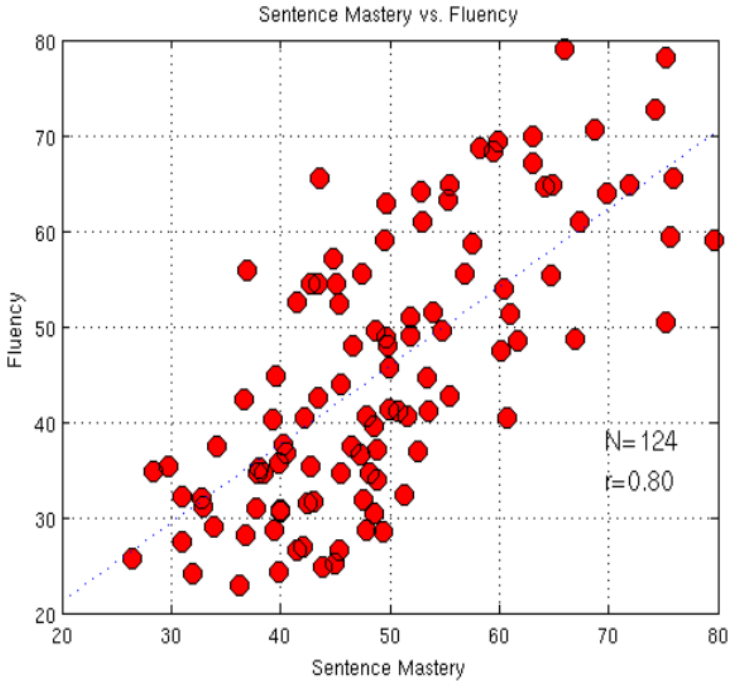

Figure 4. Sentence Mastery vs. Fluency scores for the validation sample (r = 0.80)

8.2.5 Machine Accuracy

Another analysis for internal quality of the test involved comparing scores from preliminary E^Pro, which uses automated language processing technologies, versus careful human judgments from expert raters.

Table 14 presents Pearson Product-Moment correlations between machine scores and human scores, when both methods are applied to the same performances on the same written responses. The candidate sample is the same set of the 124 validation candidates that was used in the reliability and subscore analyses. The human scores in Table 14 were calculated from a single human judgment, which means that the correlation coefficients are conservative (higher coefficients can be obtained with multiple human ratings).

Table 14. Correlation Coefficients between Human and Machine Scoring of Writing Profile Responses (n = 124).

Score	Correlation
Writing Profile	0.98
Grammar	0.99
Word Choice	0.98
Organization	0.90
Voice & Tone	0.91
Reading	0.96

The correlations presented in Table 14 suggest that the Writing Profile scores of the preliminary E^Pro test that were produced automatically by machine yielded scores that closely corresponded with human ratings. Among the subscores, the human-machine relation is closer for the linguistic (Grammar and Word Choice) and content (Reading) aspects of written language than for the rhetoric aspect (Organization and Voice & Tone), but the relation is close for all five analytic subscores.

Table 15 presents Pearson Product-Moment correlations for spoken responses. The correlations suggest that the Speaking Profile scores of the preliminary E^Pro test by machine yielded scores that closely corresponded with human ratings. Among the subscores, the human-machine relation is closer for Listening and the content aspects of spoken language (Sentence Mastery and Vocabulary) than for the manner-of-speaking subscores (Fluency and Pronunciation), but the relation is close for all five analytic subscores.

Table 15. Correlations between Human and Machine Scoring of Speaking Profile Responses (n = 124).

Score	Correlation
Speaking Profile	0.95
Sentence Mastery	0.93
Vocabulary	0.95
Fluency	0.85
Pronunciation	0.84
Listening	0.96

Both Tables 14 and 15 show that at the profile score level, machine-generated scores are virtually indistinguishable from scoring that is done by careful human transcriptions and multiple independent human judgments.

8.2.6 Differentiation among Known Populations

The next validity analysis examined whether or not scores of the preliminary E^Pro reflect expected differences between native English speakers and English language learners. Writing Profile scores from 400 native speakers and 1709 non-native speakers representing a range of native languages were compared. Figure 4 presents cumulative distributions of Writing Profile scores for the native and non-native speakers. Note that the range of scores displayed in this figure is from 10 through 90, whereas the preliminary scores were reported on a scale from 20 to 80. Scores outside the 20 to 80 range are

deemed to have saturated the intended measurement range of the test and are therefore reported as 20 or 80.

The results show that native speakers of English consistently obtain high Writing Profile scores. Fewer than 5% of the native sample scored below 70, which was mainly due to performance in Email Writing (i.e. rhetorical written skills rather than language skills). Learners of English as a second or foreign language, on the other hand, are distributed over a wide range of scores. Note also that only 10% of the non-natives scored above 70. In sum, the Writing Profile scores show effective separation between native and non-native candidates.

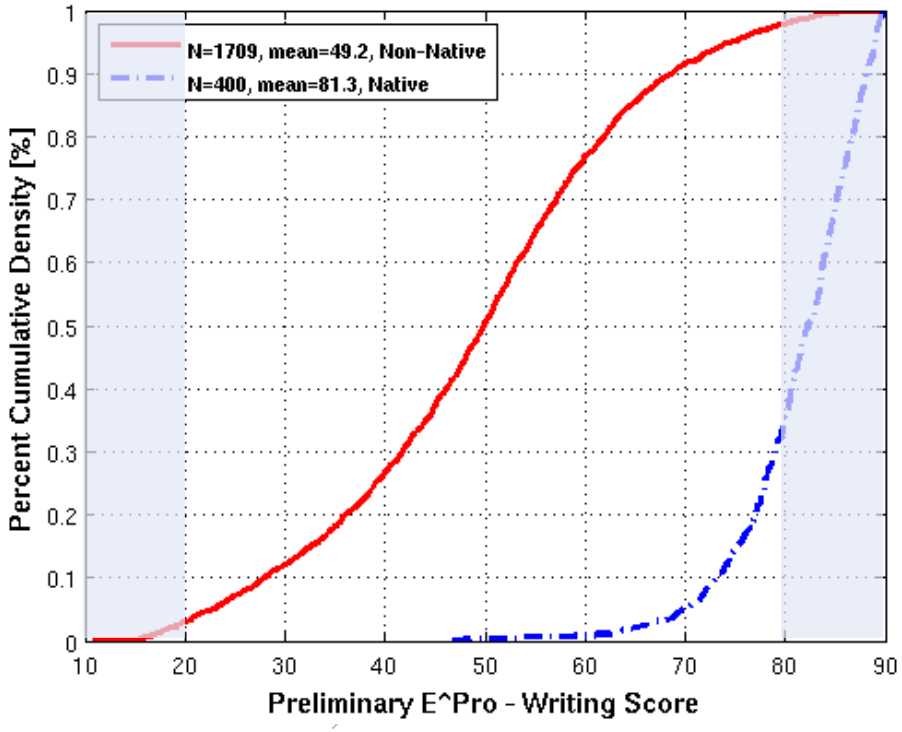

Figure 5. Cumulative density functions of Preliminary E^Pro Writing Profile scores for the native and non-native groups (native n=400 and non-native n=1709).

Similarly, the expected score differences between native English speakers and English language learners were examined for the Speaking Profile scores. As has been shown for the Writing Scores in Figure 5, Speaking Profile scores from learners should also distribute over the score range according to their spoken English ability, whereas the native speakers should receive high Speaking Profile scores.

Overall Speaking Profile scores from 28 native speakers and 987 non-native speakers representing a range of native languages were compared. Figure 6 presents the score distributions of Speaking Profile scores for the native and non-native speakers in the form of histograms. The results show that native speakers of English consistently obtain high Speaking Profile scores (in red). All native test-takers scores fall into the last score bin of 76-80. On the other hand, learners of English as a second or foreign language are normally distributed over a wide range of scores. As with the Writing Profile scores, Speaking Profile scores show effective separation between native and non-native candidates.

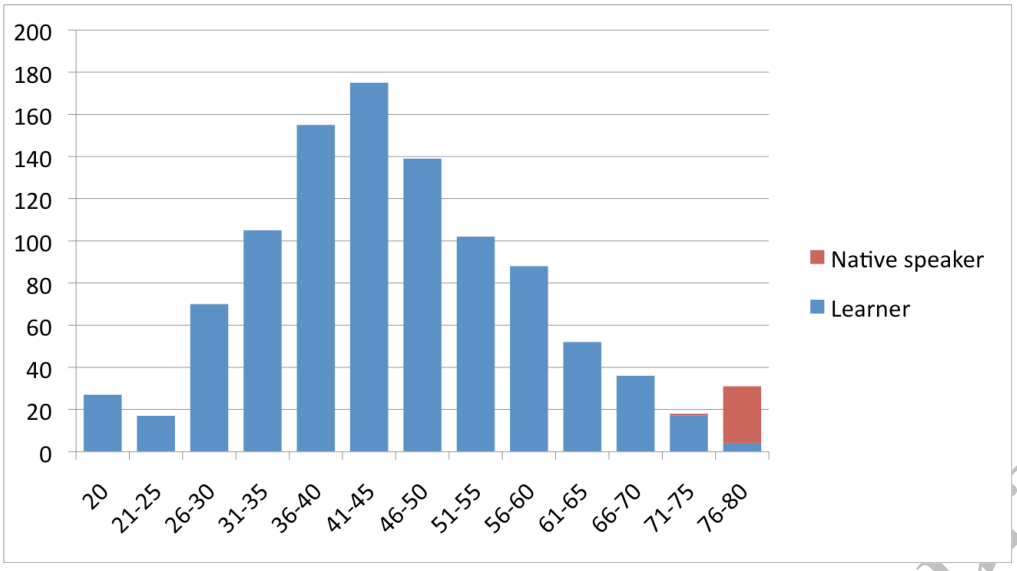

Figure 6. Histograms of Preliminary E^Pro - Speaking Profile scores for the native and non-native groups (native n=28 and non-native n=987).

8.3 Concurrent Validity

One important goal of the validity studies is to understand how E^Pro scores relate to other measures of English proficiency. Since the E^Pro exam has an emphasis on workplace English, it would be most sensible to explore a relationship with another well-known workplace English test. For this reason, a study was undertaken to compare the automatically derived scores of the preliminary E^Pro with the Test of English for International Communication (TOEIC). In addition, another study was undertaken to identify the relationship between the E^Pro Speaking Profile scores of the preliminary E^Pro and the well-established Versant English Test.

8.3.1 Preliminary E^Pro and TOEIC

The TOEIC Listening and Reading test was used as a concurrent validation. The TOEIC Listening and Reading test is claimed to measure "a non-native speaker's listening and reading skills in English as these skills are used in the workplace. The test was developed about 30 years ago as a measure of receptive language skill and has been widely accepted and used worldwide." (Liao, Qu, & Morgan, 2010). The Listening and Reading subscores are both reported in the range of 5 to 495 for a total score between 10 and 990.

Method

The study was conducted between November 2009 and February 2010. The participants were 28 Japanese and 27 South Koreans who represented a mix of full-time students and working professionals. Of the 55 participants, 26 were male and 29 female with a mean age of 24. The participants were recruited by agents in Japan and Korea acting on Pearson's behalf (a university professor and two business professionals).

The participants took both the Speaking Profile and Writing Profile versions of the preliminary E^Pro as well as TOEIC tests, with a gap between sittings of no less than 30 days. All participants were first asked to take shorter versions of the preliminary E^Pro Writing Profile and E^Pro Speaking Profile as demo tests so their resulting performance would more closely relate to their proficiency levels, rather than reflect their unfamiliarity with the E^Pro exams. They took their tests individually at their home,

school, or workplace. The TOEIC tests were administered during the official test administrations. No institutional TOEIC tests were used.

Results

The inter-correlation matrix between the subscores of each test is given in Table 16. The values across all subscores are at or above r=0.68. Not surprisingly, the highest correlation coefficients (0.96 and 0.91) exist between subscores (or profiles) and the overall scores for the same test. This is true for both the preliminary E^Pro Writing Profile and TOEIC. Since the TOEIC Listening and Reading test includes more listening items than the preliminary E^Pro Writing Profile, the overall scores from the preliminary E^Pro Writing Profile and E^Pro Speaking Profile were combined. The correlation between the preliminary E^Pro total and the TOEIC total was r=0.78. Though the sample size is small, these matrixes (below) show an expected pattern of relationships among the subscores of the tests, bearing in mind that they all relate to English language ability but assess different dimensions of that ability.

Table 16. Pearson Correlation Coefficients for Preliminary E^Pro Writing Profile and TOEIC (n=55).

	TOEIC Reading	TOEIC Listening	TOEIC Total	E^Pro Writing Profile
TOEIC Reading	-			
TOEIC Listening	0.84	-		
TOEIC Total	0.96	0.96	-	
E^Pro - Writing	0.70	0.68	0.72	-
E^Pro Total[2]	0.75	0.76	0.78	0.91

The preliminary E^Pro Writing Profile score and TOEIC Total correlated moderately at r=0.72, as shown in Figure 7, indicating that there is general English ability as a covariance, but that these tests measure different aspects of language performance (i.e., different test constructs). The preliminary E^Pro Writing Profile score correlated higher with TOEIC Reading (r=0.70) than with TOEIC Listening (r=0.68), which is expected because more content is presented through reading than listening in the preliminary E^Pro Writing Profile test.

[2] The E^Pro Total score represents an average of the overall scores from the separate modules, E^Pro - Speaking and E^Pro - Writing.

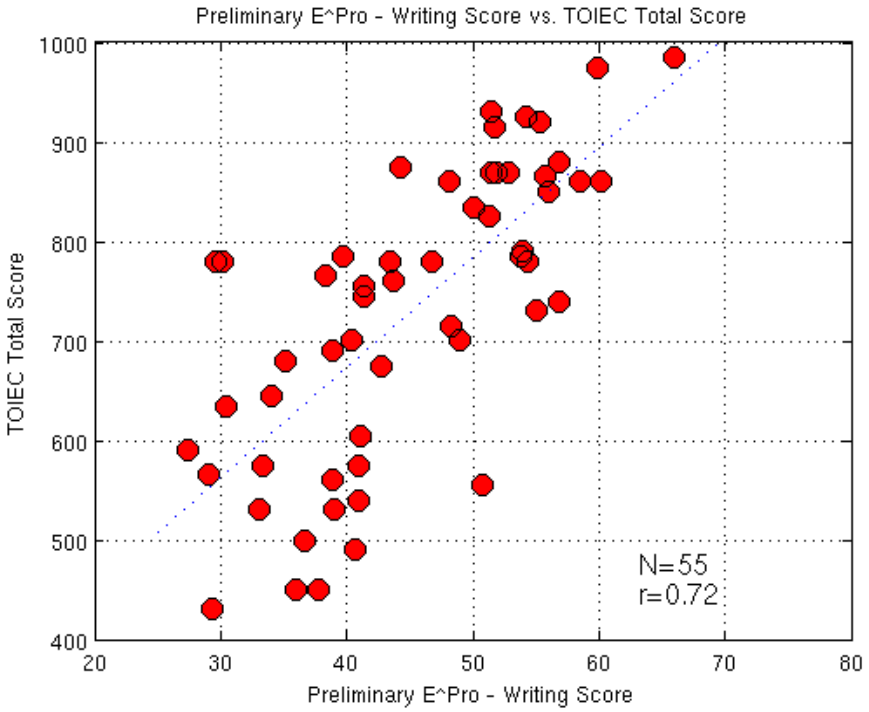

Figure 7. Scatterplot showing the relationship between Preliminary E^Pro Writing Profile and TOEIC (n=55).

For the preliminary E^Pro Speaking Profile, the inter-correlation matrix between the subscores of each test is given in Table 17. The values across all subscores are at or above r=0.67. Not surprisingly, the highest correlation coefficients (0.96 and 0.91) exist between subscores (or modules) and the overall scores for the same test. This is true for both the preliminary E^Pro Speaking Profile score and TOEIC. When the overall scores from the preliminary E^Pro Writing Profile and E^Pro Speaking Profile are combined, the correlation between the E^Pro total and the TOEIC total is r=0.78.

Table 17. Pearson Correlation Coefficients for Preliminary E^Pro - Speaking Profile and TOEIC (n=55).

	TOEIC Reading	TOEIC Listening	TOEIC Total	E^Pro - Speaking
TOEIC Reading	-			
TOEIC Listening	0.84	-		
TOEIC Total	0.96	0.96	-	
E^Pro - Speaking	0.67	0.71	0.72	-
E^Pro Total	0.75	0.76	0.78	0.91

The preliminary E^Pro Speaking Profile score and TOEIC Total correlated moderately at r=0.72, as shown in Figure 8, indicating that there is general English ability as a covariance, but that these tests measure different aspects of language performance. The preliminary E^Pro Speaking Profile score correlated higher with TOEIC Listening (r=0.71) than with TOEIC Reading (r=0.67), which is expected because E^Pro Speaking Profile is designed to measure listening and speaking abilities rather than reading ability.

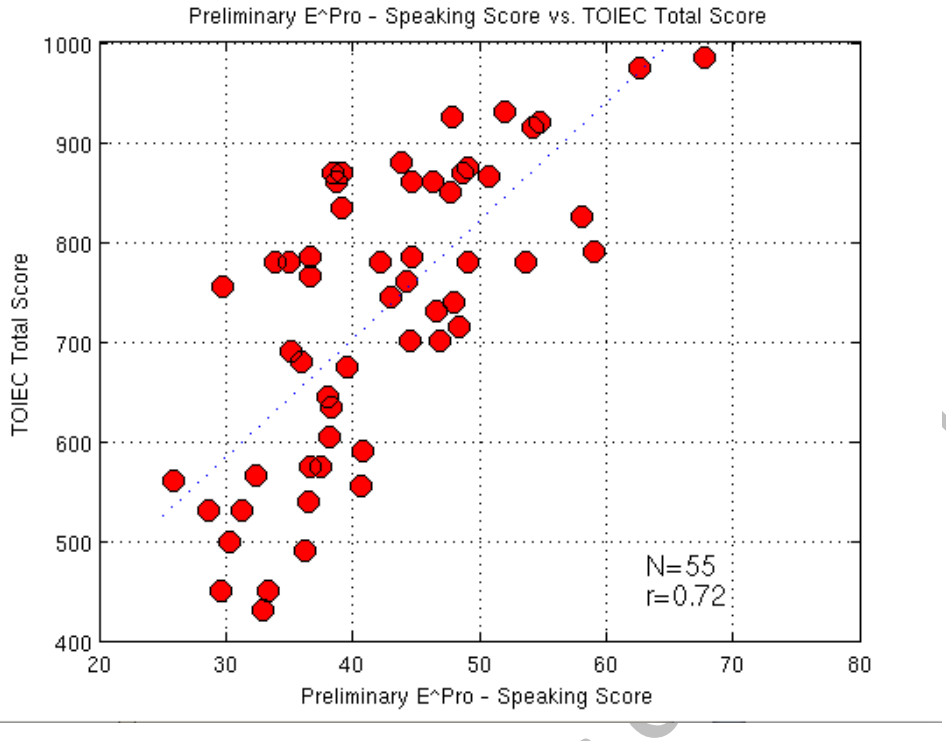

Figure 8. Scatterplot showing the relationship between Preliminary E^Pro Speaking Profile and TOEIC (n=55).

8.3.2 E^Pro Speaking Profile and Versant English Test

A study was conducted to explore the relation between the preliminary E^Pro Speaking Profile scores and the Versant English Test (VET) scores. Both tests are designed to measure facility in spoken English. Empirical evidence has demonstrated that the Versant English Test is a valid tool to assess spoken English. If there is a close relation E^Pro Speaking Profile and the Versant English Test, it then follows that E^Pro Speaking Profile also measures what it claims to measure – facility in spoken English.

The analysis involved the validation set of 124 candidates taking the preliminary E^Pro Speaking Profile and extracting their performances on the parts of the test that share similarities with the Versant English Test (i.e. Repeats, Shorts Answer Questions, Sentence Builds, and Story Retells). These responses were processed through the automated scoring algorithms for the Versant English Test and the scores were compared to the preliminary E^Pro Speaking Profile scores, as shown in Figure 9.

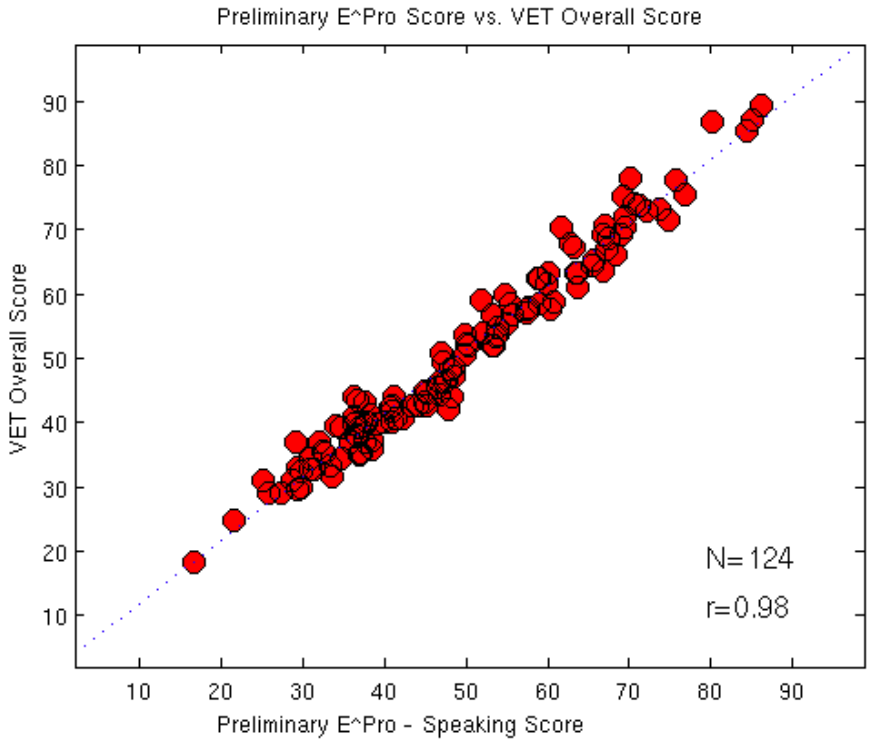

Figure 9. Scatterplot showing the relation between Preliminary E^Pro Speaking Profile scores and Versant English Test scores.

As can be seen, there is a strong relation between the two sets of scores (r=0.98). This strong relation between the E^Pro Speaking Profile and the Versant English Test supports the claim that E^Pro Speaking Profile measures the intended construct, i.e., facility with spoken English.

Although the relation between the two tests is strong when investigated at the entire sample level, there are clear individual differences between candidates, as shown in Table 18. Preliminary E^Pro Speaking Profile and Versant English Test scores correspond to one another at essentially a one-to-one correspondence (e.g. 47 on one test is equal to 47 on the other test), but the scores of individuals vary depending on the personal skill set.

Table 18. Score comparison between Preliminary E^Pro Speaking Profile and the Versant English Test.

Proportion of candidates (n=124)	Score point difference between E^Pro Speaking Profile and VET
1 %	≥ 9 points difference
10 %	≥ 6 points difference
31 %	≥ 4 points difference
58 %	≤ 3 points difference

This difference in scoring between the preliminary E^Pro Speaking Profile and the Versant English Test is largely due to the impact of the Listening subscore in E^Pro Speaking Profile. The correlation coefficient of Listening with the other four subscores of the test combined (Sentence Mastery, Vocabulary, Fluency and Pronunciation) was r=0.84. This high correlation reveals that Listening shares common variance with the other subscores but also contributes unique information about the candidate's spoken English ability when combined as the Speaking Profile scores.

8.4 Benchmarking to Common European Framework of Reference

8.4.1 E^Pro Writing Profile and CEFR Level Estimates

In order to identify the correspondence between scores on the preliminary E^Pro Writing Profile and CEFR, a standard-setting procedure was conducted following the guidelines of the Manual for Relating Language Examinations to the Common European Framework of Reference (Council of Europe, 2001). The goal was to identify minimum scores (cut scores) on the preliminary E^Pro Writing Profile that map to the A1 through C2 proficiency levels of the CEFR.

Method

A set of analytic descriptors containing six levels was developed from the CEFR scales, corresponding to CEFR levels A1, A2, B1, B2, C1, and C2. Six English language testing experts were recruited as expert judges. They were instructed to utilize the CEFR descriptors to grade holistically, and choose the CEFR level that best fit each response. A response set of written samples was created using the following procedure: 240 candidates who took the preliminary E^Pro Writing Profile were selected via stratified random sampling. This sampling technique was used to assure that the response set contained written samples from a wide variety of language backgrounds and equally distributed proficiency levels, approximately 40 per CEFR level. The candidates came from China, Costa Rica, France, Germany, India, Iran, Japan, Korea, Mexico, the Netherlands, Russia, Spain, Taiwan, Thailand, and the United States.

Eleven of the candidates were excluded from analysis either before or after the rating process due to incomplete data (most or all responses were blank), leaving 229 individual candidates in the response set. Each candidate contributed a total of five written responses from two tasks: three Passage Reconstruction responses and two Email Writing responses. The response set therefore consisted of 1145 written samples: 687 Passage Reconstruction responses and 458 Email Writing responses.

Results

Raters demonstrated a high level of consistency with one another in their assigned scores (r=0.98). This high level of inter-rater reliability demonstrates that candidates can be consistently classified into CEFR levels based on performances elicited by these tasks. The CEFR ratings from the six raters and the preliminary E^Pro Writing Profile for each candidate were entered into a Rasch model to produce an ability estimate for each candidate on a common logit scale. Initial CEFR boundaries were then estimated from Rasch ability estimates, as shown in Table 19.

Table 19. CEFR Score Boundaries as Logits from a Rasch Model.

Facetstep	CEFR Level	Expectation Measure at CEFR Boundary (Logits)
1	A1	-4.43
2	A2	-2.45
3	B1	-0.68
4	B2	0.88
5	C1	2.39
6	C2	4.22

Candidates' E^Pro Writing Profile scores were then lined up next to their CEFR-based ability estimates to establish the score boundaries. When comparing the aggregated expert judgments with the preliminary E^Pro Writing Profile scores to establish a CEFR Level, 68% of candidates are correctly classified and 99% of candidates are classified correctly or one level away. Table 20 below provides the final mapping between the two scales.

Table 20. Mapping of CEFR Levels with Preliminary E^Pro Writing Profile Scores.

CEFR Level	E^Pro Writing Profile Score Range
A1	20-29
A2	30-43
B1	44-53
B2	54-66
C1	67-76
C2	77-80

Figure 10 plots the relation between each candidate's E^Pro Writing Profile score (shown on the x-axis) and their *CEFR ability estimate* in logits as estimated from the judgments of the six panelists (shown on the y-axis). The figure also shows the original Rasch-based CEFR boundaries (horizontal dotted lines) and the slightly adjusted boundaries (vertical dotted lines).

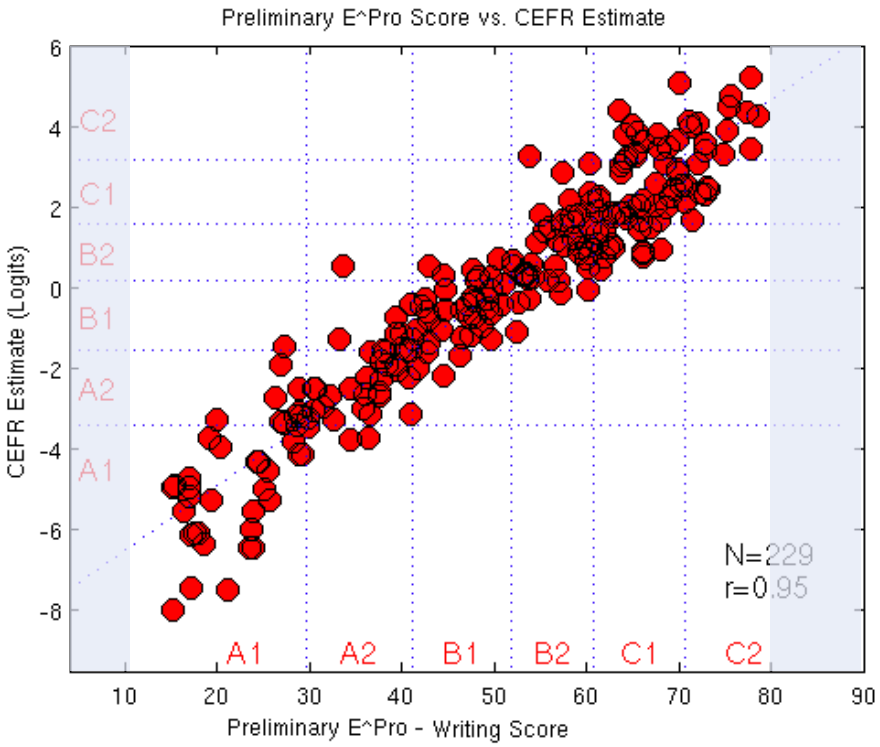

Figure 10. Scatterplot showing Rasch-based CEFR ability estimates as derived from human judgments and Preliminary E^Pro Writing Profile scores.

The Pearson correlation coefficients for Preliminary E^Pro Writing Profile scores and CEFR estimates is 0.95, revealing that Preliminary E^Pro Writing Profile yields test scores which are highly consistent with judges' evaluation of written performance using the CEFR scales.

The raters' CEFR ratings were based on two tasks (Email Writing and Passage Reconstruction) which elicit linguistic, content and rhetorical skills. However, it is important to note that the E^Pro Writing Profile score is derived not only from performance on these two tasks, but also on Sentence Completion and Dictation which assess linguistic skills more reliably. Therefore, some error in CEFR classification is to be expected when individuals have substantially different linguistic skills than content and rhetorical skills.

8.4.2 E^Pro Speaking Profile and CEFR Level Estimates

Because of the high correlation (r=0.98) between E^Pro Speaking Profile and the Versant English Test (see section 8.3.2), the results from a previous study mapping Versant English scores onto the CEFR levels have been applied to E^Pro Speaking Profile. That is, the established Versant English score ranges aligned with the CEFR levels have been used for E^Pro Speaking Profile, as shown in Table 21. The method used to create the mappings is described in the *Can-Do Guide*. Please contact Pearson for this report.

Table 21. Mapping of CEFR Levels with E^Pro Speaking Scores.

CEFR Level	E^Pro Speaking Profile Score Range
<A1	20-25
A1	26-35
A2	36-46
B1	47-57
B2	58-68
C1	69-78
C2	79-80

9. Conclusion

This report has provided details of the test development process and validity evidence for the English for Professionals Exam. The information is provided for test users to make an informed interpretive judgment as to whether test scores would be valid for their purposes. The test development process is documented and adheres to sound theoretical principles and test development ethics from the field of applied linguistics and language testing:

- the items were written to specifications and were subjected to a rigorous procedure of qualitative review and psychometric analysis before being deployed to the item pool;
- the content was selected from both pedagogic and authentic material;
- the test has a well-defined construct that is represented in the cognitive demands of the tasks;
- the scores, item weights and scoring logic are explained;
- the items were widely field tested on a representative sample of candidates.

This report provides empirical evidence demonstrating that the preliminary E^Pro scores are structurally reliable indications of candidate ability in written English and are suitable for high-stakes decision-making.

10. About the Company

Pearson: Pearson's Knowledge Technologies group and Corporation, the creator of the Versant tests, were combined in January, 2008. The Versant line of tests is the first to leverage a completely automated method for assessing spoken language.

Versant Testing Technology: The Versant automated testing system was developed to apply advanced speech recognition techniques and data collection to the evaluation of language skills. The system includes automatic telephone and computer reply procedures, dedicated speech recognizers, speech analyzers, databanks for digital storage of speech samples, and score report generators linked to the Internet. The English for Professionals Exam is the result of years of research in statistical modeling, linguistics, testing theory and speech recognition. The Versant patented technologies are applied to Pearson's own language tests such as the Versant series and also to customized tests. Sample projects include assessment of spoken English, assessment of spoken aviation English, children's reading assessment, adult literacy assessment, and collections and human rating of spoken language samples.

Pearson's Policy: Pearson is committed to the best practices in the development, use, and administration of language tests. Each Pearson employee strives to achieve the highest standards in test publishing and test practice. As applicable, Pearson follows the guidelines propounded in the Standards for Educational and Psychological Testing, and the Code of Professional Responsibilities in Educational Measurement. A copy of the Standards for Educational and Psychological Testing is available to every employee for reference.

Research at Pearson: In close cooperation with international experts, Pearson conducts ongoing research aimed at gathering substantial evidence for the validity, reliability, and practicality of its current products and at investigating new applications for Versant technology. Research results are published in international journals and made available through the Versant test website (http://www.VersantTest.com).

11. References

Alderson, J. C. (2000). Assessing reading. Cambridge: Cambridge University Press.

Bachman, L.F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Bachman, L.F. & Palmer, A.S. (1996). Language testing in practice. Oxford: Oxford University Press.

Buck, G. (2001). Assessing listening. Cambridge: Cambridge University Press.

Bull, M & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In R.H. Mannell & J. Robert-Ribes (Eds.), Proceedings of the 5th International Conference on Spoken Language Processing. Canberra, Australia: Australian Speech Science and Technology Association.

Carroll, J.B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. Testing. Washington, DC: Center for Applied Linguistics.

Carroll, J.B. (1986). Second language. In R.F. Dillon & R.J. Sternberg (Eds.), Cognition and Instructions. Orlando FL: Academic Press.

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

Cutler, A. (2003). Lexical access. In L. Nadel (Ed.), Encyclopedia of Cognitive Science. Vol. 2, Epilepsy – Mental imagery, philosophical issues about. London: Nature Publishing Group, 858-864.

Godfrey, J.J. & Holliman, E. (1997). Switchboard-1 Release 2. LDC Catalog No.: LCD97S62. http://www.ldc.upenn.edu.

Goodman, K. (1969). Analysis of oral reading miscues: Applied psycholinguistics. In F. Gollasch (Ed.) Language and literacy: The selected writings of Kenneth Goodman (pp. 123–134). Vol. I. Boston: Routledge & Kegan Paul.

Gough, P.B., Ehri, L.C., and Treiman, R. (1992). Reading acquisition. Hillsdale, NJ: Erlbaum.

Grabe, W., and Kaplan, R.C. (1996). Theory and practice of writing. New York: Longman.

Jescheniak, J.D., Hahne, A. & Schriefers, H.J. (2003). Information flow in the mental lexicon during speech planning: evidence from event-related brain potentials. Cognitive Brain Research, 15(3), 261-276.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 387-412.

Levelt, W.J.M. (1989). Speaking: From intention to articulation. Cambridge, MA: MIT Press.

Levelt, W.J.M. (2001). Spoken word production: A theory of lexical access. PNAS, 98(23), 13464-13471.

Liao, C-w., Qu, Y., and Morgan, R. (2010). The Relationship of Test Scores Measured by the TOEIC® Listening and Reading Test and TOEIC® Speaking and Writing Tests (TC-10-13). Retrieved from Educational Testing Service website: http://www.ets.org/research/policy_research_reports/tc-10-13

Luoma. (2003). Assessing speaking. Cambridge: Cambridge University Press.

McLaughlin, G.H. (1969). SMOG grading: A new readability formula. Journal of Reading, 12(8), 639-646.

Miller, G.A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior, 2, 217-228.

Oakeshott-Taylor, J. (1977). Information redundancy, and listening comprehension. In R. Dirven (ed.), Hörverständnis im Fremdsprachenunterrict. Listening comprehension in foreign language teaching. Kronberg/Ts.: Scriptor.

Oller, J. W. (1971). Dictation as a device for testing foreign language proficiency. English Language Teaching, 25(3),254-259.

Ordinate (2003). Ordinate SET-10 Can-Do Guide. Menlo Park, CA: Author.

Perry, J. (2001). Reference and reflexivity. Stanford, CA: CSLI Publications.

Segalowitz, N., Poulsen, C., and Komoda, M. (1991). Lower level components or reading skill in higher level bilinguals: Implications for reading instruction. In J.H. Hulstijn and J.F. Matter (eds.), Reading in two languages, AILA Review, Vol. 8,. Amsterdam: Free University Press, 15-30.

Sigott, G. (2004). Towards identifying the C-test construct. New York: Peter Lang.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. Language Testing, 14(2), 214-231.

Stansfield, C.W., & Kenyon, D.M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*, 129-141.

Van Turennout, M. Hagoort, P., & Brown, C.M. (1998). Brain activity during speaking: From syntax to phonology in 40 milliseconds. *Science, 280*, 572-574.