Versant™ Writing Test

Test Description and Validation Summary

Table of Contents

1. Introduction

Pearson's Versant™ Writing Test (VWT), powered by Versant technology, is an assessment instrument designed to measure how well a person can handle workplace English in written form. The VWT is intended for adults 18 years of age and older and takes about 40 minutes to complete. Because the VWT is delivered automatically by the Versant testing system, the test can be taken at any time, from any location on a computer. A human examiner is not required. The computerized scoring allows for immediate, objective, and reliable results that correspond well with traditional measures of English language proficiency.

The VWT measures *facility* in written English in the workplace context. Facility is defined as *how well a person can understand spoken or written English and respond in writing appropriately on everyday and workplace topics at a functional pace*. VWT scores provide reliable information that can be applied to placement, qualification and certification decisions by academic institutions, businesses and government agencies. The test is also appropriate for monitoring progress as well as measuring instructional outcomes. (The Versant English Test (VET) is also available if it is necessary to evaluate spoken English. For more information about the VET, please refer to *Versant English Test: Test Description and Validation Summary*.)

2. Test Description

2.1 Workplace Emphasis

The VWT 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, the VWT is intended to assess how well and how efficiently the candidate can process written English on general topics such as scheduling, commuting, and training that are commonly found in the workplace regardless of industry or job category.

2.2 Test Design

The VWT has five automatically scored tasks: Typing, Sentence Completion, Dictation, Passage Reconstruction, and Email Writing. These tasks provide multiple, fully independent measures that underlie facility in written English, including sentence comprehension and construction, passive and active vocabulary use, and appropriateness and accuracy of writing. Because more than one task contributes to many of the subscores, the use of multiple item types strengthens score reliability.

The VWT score report is comprised of an Overall score and five subscores: Grammar, Vocabulary, Voice and Tone, Organization, and Reading Comprehension.

The Overall score is a weighted average of the five subscores. Together, these scores describe the candidate's facility in written English in everyday and workplace contexts. As supplemental information, Typing Speed and Typing Accuracy are also reported on the score report.

Once a candidate has completed a test, the responses are sent to a remote server, from which the Versant testing system automatically analyzes them and posts scores at www.VersantTest.com usually within minutes of completing the test. Test administrators and score users can view and print out test results from ScoreKeeper.

2.3 Test Administration

The VWT is administered via Versant for Web (VfW), a browser-based system. It is available in both an on-line and off-line mode. The VWT can be taken at any time, from any location. Automated administration eliminates the need for a human examiner. However, depending on the test score use, a proctor may be necessary to verify the candidate's identity and/or to ensure that the test is taken under exam conditions. The VWT can also be administered via Pearson's Computer Delivered Test (CDT) software, which "locks down" the computer to prevent web-browsing, consulting files on the local hard drive, copying or pasting, etc.

The candidate must use a microphone headset to take the VWT in order to guarantee a consistent sound quality of both test content and responses. When the test is launched, the candidate is prompted to enter a unique Test Identification Number (TIN) using the keyboard.

During the test administration, an examiner's voice guides the candidate through the test, explains the test tasks, and gives examples. Candidates interact with the test system in English, typing their responses using a keyboard. When the test is finished, the candidate clicks a button labeled, "End Test".

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. If candidates do not finish a response in the allotted time, their work is saved automatically and the next item begins. If the candidate finishes before the allotted time has run out, they can click a button labeled "Next" to move on to the next item.

2.4 Test Format

The following subsections provide brief descriptions of the tasks and the abilities required to respond to the items in each of the five parts of the VWT.

Part A: Typing

The VWT includes a typing speed and accuracy task which is not included in the actual test scores. In this task, candidates see a passage at the top of the computer screen and have 60 seconds to type the passage exactly as it appears into a box at the bottom of the screen. All passages deal with different aspects of typical business topics or activities. The passages are relatively simple in structure and vocabulary and range in length from 120 to 130 words. The SMOG Readability Index 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 10 and 12, which is at a high school level and can be easily typed by most educated English speakers with adequate typing skills.

Example:

> Whenever you have a fantastic idea, you should always write it down. If you don't, it is quite possible that you will forget about it. Many creative people have a pen and paper close at hand at all times. That way, whenever an interesting thought comes to them, they are prepared to write it down. Later on, when they have time, they sit down and read through their list of ideas.
>
> You can benefit from this practice, too. Keeping a notebook full of thoughts is a great way of understanding yourself better, because it tells you how you think. It allows you to return to an interesting idea when you have the opportunity to do so. You might find that you've created something that can change the world forever.

This task has several functions. First, since typing is a familiar task to most candidates, it is a comfortable introduction to the interactive mode of the VWT as a whole. Second, it allows candidates to familiarize themselves with the keyboard. Third, it measures the candidate's typing speed and accuracy. The VWT assumes a basic competence in typing for every candidate. Since it is important to disambiguate candidates' typing skills from their written English proficiency, it is recommended that test administrators review each candidate's typing score. If typing speed is below 12 words per minute, and/or accuracy is below 90%, then it is likely that this candidate's written English proficiency was not properly measured due to poor typing skills. The test administrator should take this into account when interpreting test scores.

Part B: Sentence Completion

In the Sentence Completion 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, think of an appropriate word, and type the word in the text box provided to complete the sentence. Sentences range in length from 4 to 30 words, and the missing words are in different positions in sentences and are of various parts of speech (e.g., noun, verb, adjective, adverb).

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 (in some cases 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 reflects the candidate's mastery of vocabulary in use.

Part C: 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 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), which refers to the range of contextually-influenced choices made by language users. Proficient listeners tend to understand and remember the content of a message, but 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).

Dictation is a good test of comprehension, language processing, and writing ability. As the sentences increase in length and complexity, the task becomes increasingly difficult for candidates who are less 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).

Part D: Passage Reconstruction

Passage Reconstruction is similar to a task known as free recall, or immediate recall. Candidates are asked 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 and 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.

In order to perform this task, the candidate must read the passage presented, understand the content, and hold it in memory long enough to reconstruct the passage in writing. Individual candidates may naturally employ various strategies when performing this task. Reconstruction may be more or less 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.

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.

(*Email*) 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.

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 first language then the main ability assessed is reading comprehension, but when the passage is reconstructed in the target language (in this case, English), then it is more an integrated test of both reading and writing (Alderson, 2000).

Part E: Email Writing

In the Email Writing 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 are asked to 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 which are off-topic are assigned the lowest possible score.

Each email situation contains several elements:

- the setting or place 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 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 suggestions.

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, Vocabulary, Voice & Tone, Organization, and Reading Comprehension subscores.

2.5 Number of Items

In the administration of the Versant Writing Test, the testing system presents approximately 43 items in five separate sections to each candidate. The items are drawn at random from a large item pool. This means that most or all items are different from one test administration to the next. Proprietary algorithms are used by the testing system to select from the item pool – the algorithms take into consideration, among other things, an item's difficulty level and similarity to other presented items. Table 1 shows the approximate number of items presented in each section. The exact number of items in each test may change from time to time as new, unscored items are added to and removed from the test. The responses to the unscored items do not impact the candidates' scores nor do they impact the test experience. The responses are used to build scoring models for new items, which allows Pearson to add new content to the test in order to keep the item bank secure and up-to-date.

Table 1. Approximate number of items presented per task

Task	Approximate Number of Items
A. Typing	1
B. Sentence Completion	20
C. Dictation	16
D. Passage Reconstruction	4
E. Email Writing	2
Total	**43**

3. Test Construct

3.1 Facility in Written English

For any language test, it is essential to define the test construct as explicitly as possible (Bachman, 1990; Bachman & Palmer, 1996). The VWT is designed to measure a candidate's *facility* in written English in the workplace context, which is *how well the person can understand spoken or written English and respond in writing appropriately on everyday and workplace topics at a functional pace*.

The constructs that can be observed in the candidate's performances in the VWT 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 receptive modalities (listening and reading) are used in the test. Dictation exposes candidates to spoken English, and the remaining sections present written English that candidates must read and comprehend within given time limits.

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 something that differentiates first language readers from even highly proficient second language readers (Segalowitz, Poulsen, & Komoda, 1991). First language 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 or written language than users of lower proficiency do, and they 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, Ehri, & Trieman, 1992). Alderson (2000) 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, the VWT 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 VWT 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 the VWT 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 and tone. 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 second language (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; and 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 in the workplace is email, the VWT 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 the VWT 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 VWT subscores of Grammar and Vocabulary (linguistic knowledge), Organization (discourse knowledge), and Voice & Tone (sociolinguistic knowledge).

Table 2. Taxonomy of Language Knowledge (adapted and simplified from Grabe and Kaplan, 1996)

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 2, 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 the VWT 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.

3.2 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.

Except for the Email Writing task, all 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 Dictation, Sentence Completion, and Passage Reconstruction tasks present context-independent material while 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.

3.3 The Role of Memory

Some measures of automaticity can be misconstrued as memory tests. Since some VWT tasks involve repeating long sentences, holding sentence in memory in order to type them, or re-assembling paragraphs from memory, it may seem that these tasks are unduly influenced by general memory

performance. Note that every Dictation and Passage Reconstruction item on the test was presented to a sample of educated native speakers of English. If memory, as such, were an overriding component of performance on the VWT tasks, then native English speakers should show greater performance variation on these items according to the presumed range of individuals' memory spans (see Section 8.2.5 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 7.3.2, CEFR Level Estimates).

4. Content Design and Development

4.1 Vocabulary Selection

The vocabulary used in the test 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

The VWT items were drafted by trained item writers. All item 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 90% of a reference sample of educated native speakers of English.

4.3 Item Prompt Recording

4.3.1 Voice Distribution

Two native speakers (one male and one female) were selected for recording the spoken prompts in the Dictation section. A professional male voice recorded the examiner prompts for the test.

4.3.2 Recording Review

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

5. Score Reporting

5.1 Scoring and Weighting

The VWT score report is comprised of an Overall score and five subscores (Grammar, Vocabulary, Organization, Voice & Tone, and Reading Comprehension).

Overall: The Overall score of the test represents the ability to understand English and respond appropriately in writing at a functional pace for everyday and workplace purposes. Scores are based on a weighted combination of the five subscores. Scores are reported in the range from 10 to 90 on Pearson's Global Scale of English (GSE). The corresponding Common European Framework of Reference for Languages (CEFR) level is also displayed.

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.

Vocabulary: Vocabulary 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.

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.

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.

Reading Comprehension: Reading 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.

Table 3 shows how the five subscores are weighted to achieve an Overall score.

Table 3. Subscore weighting in relation to VWT Overall score

Subscore	Weight
Grammar	30 %
Vocabulary	30 %
Organization	10 %
Voice & Tone	10 %
Reading Comprehension	20 %
Overall Score	**100 %**

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).

The linguistic aspect is informed by the Grammar and Vocabulary subscores. Combined, these two dimensions account for 60% 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 Comprehension subscore. This accounts for 20% of the Overall 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 Overall 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.

In the VWT scoring logic, the linguistic, content, and rhetoric aspects are weighted 60, 20, and 20, respectively, to reflect their importance for successful written communication.

5.2 Score Use

Once a candidate has completed a test, the candidate's responses are sent to a remote server, from which the Versant testing system analyzes them and posts scores at www.VersantTest.com. Test administrators and score users can view and print out the test results from ScoreKeeper.

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

The VWT score scale covers a wide range of abilities in written English communication. It is up to score users to decide what VWT score can be regarded as a minimum requirement in their context (a "cut score"). Score users may wish to base their selection of an appropriate criterion score on their own localized research. Pearson can provide assistance in helping organizations to arrive at data-based criterion scores.

6. Data Resources for Scoring Development

6.1 Data Collection

Both native speakers and non-native speakers of English were recruited as participants to take a prototype data-collection version of the VWT. The purposes of this field testing were 1) to validate operation of the test items with both native and non-native speakers, 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 English samples to develop automatic scoring models for written English. The description of participants in the field testing whose responses were used to develop automated scoring models is presented in Table 4.

Table 4. Description of participants in the scoring model development (*N* = 1,768)

	Native speaker of English	Non-native speaker of English
Number of Participants	*n* = 73	*n* = 1,695
Gender	Male: *n* = 23 (31%) Female: *n* = 46 (63%) Unknown: *n* = 4 (6%)	Male: *n* = 746 (44%) Female: *n* = 830 (49%) Unknown: *n* = 119 (7%)
Age	Range: 20 to 73 Average: 35.6	Range: 19 to 67 Average: 28.0
First Language	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

6.1.1 Native Speakers

A total of 73 educated adult native speakers of English 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 and female ratio was 31:63.

While the VWT is specifically designed for non-native speakers of English, 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 Non-native Speakers

A total of 1,695 non-native candidates were recruited from various countries representing both university students and working professionals.

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 and female ratio was 44:49 with 7% of the candidates being unreported. The mean candidate age was 28.

6.2 Expert Human Rating

During the field test of the VWT, more than 50,000 responses were collected from native speakers and English learners. Subsets of the response data were presented to trained raters for developing the automatic scoring models.

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. Before the native speakers began rating the 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 in to a web-based rating system and evaluated the written responses to Passage Reconstruction and Email Writing items for such traits as vocabulary, grammar, organization, voice and tone, email conventions, and task completion. Rating stopped when each item had been judged by at least two raters.

7. Validation

7.1 Validity Study Design

A series of validity analyses were conducted to examine five aspects of the VWT scores. All scores, statistics, and results in the validation studies below (§7.1-7.4) use the original Versant scale of 20 to 80 rather than the GSE of 10 to 90.

Structural Validity

1. *Reliability:* whether or not the VWT is structurally reliable and assigns scores consistently
2. *Dimensionality:* whether or not the five different subscores of the VWT are sufficiently distinct
3. *Accuracy:* whether or not the automatically scored VWT scores are comparable to the scores that human listeners and raters would assign
4. *Differentiation among known populations:* whether or not VWT 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 VWT scores predict the reliable information in scores of a well-established English test for a workplace context (e.g., TOEIC); and how do VWT scores correspond to the six levels of the Common European Framework of Reference (CEFR).

7.1.1 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

written 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 VWT. A total of five native speakers were included in the validation dataset. Table 5 below summarizes the demographic information validation participants whose responses were used to estimate the reliability and validate the scoring model.

Table 5. Description of participants in the scoring model validation (*N*= 124).

Number of Participants	124 (including 5 native speakers)
Gender	Male: *n* = 55 (44%) Female: *n* = 69 (56%)
Age	Range: 19 to 66 Average: 30.4
First Language	Arabic, Chinese, English, Filipino, French, German, Hindi, Italian, Japanese, Korean, Malayalam, Russian, Spanish, Tagalog, Tamil, Telugu, Visayan

7.2 Structural Validity

To understand the consistency and accuracy of the VWT Overall scores and the distinctness of the subscores, the following were examined: descriptive statistics of the validation sample, the standard error of measurement of the VWT Overall score; the reliability of the VWT (split-half reliability); the correlations between the VWT Overall score and its subscores, and between pairs of subscores; comparison of machine-generated VWT scores with listener-judged scores of the same VWTs. These qualities of consistency and accuracy of the test scores are the foundation of any valid test.

7.2.1 Descriptive Statistics

Table 6 summarizes some descriptive statistics for the validation sample. The mean Overall score of the validation sample was 51.74 with a standard deviation of 15.27 (on a scale of 20-80).

Table 6. Descriptive Statistics for the Validation Dataset (*N* = 124)

Measure	Statistic
Mean	51.74
Standard Error	1.37
Median	51.55
Standard Deviation	15.27
Sample Variance	233.07
Kurtosis	-0.44
Skewness	0.06

7.2.2 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, p.183). If a candidate were to take the same test repeatedly (with no new learning taking place between tests), the standard deviation of his/her repeated test scores is denoted as the standard error of measurement. The standard error of measurement of the VWT Overall score is 2.2. In other words, if a candidate received an Overall score of 50 on the VWT and then took the test again, his or her Overall score is expected to fall between 47.8 and 52.2 on the second test.

7.2.3 Test Reliability

Score reliabilities were estimated by the split-half method. Split-half reliability was calculated for the Overall score and all 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 7.

The values in Table 7 suggest that there is sufficient information in a VWT 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[1] 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 7. Split-half reliabilities of VWT human scoring versus machine scoring (*N* = 124)

Score	Split-half Reliability for Human Scores	Split-half Reliability for Machine Scores
Overall	.93	.98
Grammar	.97	.98
Vocabulary	.89	.91
Organization	.77	.87
Voice & Tone	.79	.90
Reading Comprehension	.92	.93

[1] The possible reliability coefficient range is 0 to 1. The closer the coefficient is to 1.0, the greater the reliability is.

The reliability estimates for the Organization and Voice & Tone subscores are lower than the reliability estimates 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 .90, and inter-rater reliability for Voice & Tone was .93 at the item level (corrected for under-estimation).

7.2.4 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 8 presents the correlations among the VWT subscores and the Overall score for the same validation sample of 124 candidates, which includes five native English speakers.

Table 8. Correlation between Subscores and Overall Score on the VWT (*N* = 124)

	Grammar	Vocabulary	Organization	Voice & Tone	Reading Comprehension	Overall
Grammar	-					.96
Vocabulary	.81	-				.96
Organization	.77	.81	-			.89
Voice & Tone	.79	.83	.98	-		.91
Reading Comprehension	.91	.88	.87	.89	-	.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.

7.2.5 Machine Accuracy

An analysis for internal quality of the test involved comparing scores from the VWT, which uses automated language processing technologies, versus careful human judgments from expert raters.

Table 9 presents Pearson product-moment correlations between machine scores and human scores, when both methods are applied to the same performances on the same VWT responses. The candidate sample is the same set of 124 validation candidates that was used in the reliability and subscore analyses.

Table 9. Correlations between Human and Machine Scoring of VWT Responses (*N* = 124)

Score Type	Correlation
Overall	.98
Grammar	.99
Vocabulary	.98
Organization	.90
Voice & Tone	.91
Reading Comprehension	.96

The correlation estimates presented in Table 9 suggest that scoring a VWT by machine will yield scores that closely correspond with human ratings. Among the subscores, the human-machine relation is closer for the linguistic (Grammar and Vocabulary) and content (Reading Comprehension) aspects of written language than for the rhetoric aspect (Organization and Voice & Tone), but the relation is close for all five subscores. At the Overall score level, VWT machine-generated scores are virtually indistinguishable from scoring that is done by multiple independent human judgments.

7.2.6 Differentiation among Known Populations

The next validity analysis examined whether or not the VWT scores reflect expected differences between native English speakers and English language learners. Overall scores from a total of 400 tests completed by the native speakers and 1709 tests completed by the learners representing a range of native languages were compared. Figure 2 presents cumulative distributions of Overall scores for the native and non-native speakers. Note that the range of scores displayed in this figure is from 10 through 90 on the Versant scale (not the GSE). 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 scores on the VWT. 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 Overall scores show effective separation between native and non-native candidates.

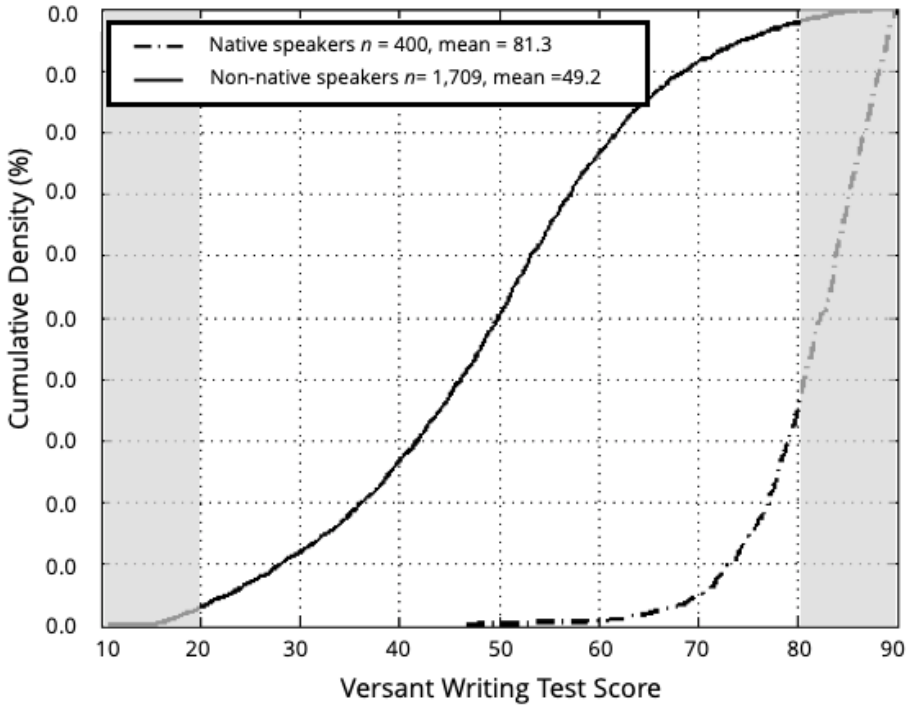

Figure 2. Cumulative density functions of VWT Overall scores for the native speakers of English and non-native speakers of English

7.3 Concurrent Validity

One important goal of the validity studies is to understand how the VWT relates to other measures of English proficiency. Since the VWT 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 VWT Overall scores with the Test of English for International Communication (TOEIC). In addition, another study was undertaken to identify the correspondence between the scores on the VWT and the six levels of the Common European Framework of Reference (CEFR).

7.3.1 VWT 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 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 VWT and TOEIC with a gap between sittings of no less than 30 days. All participants were asked to take a shorter version of the VWT as a demo test so their resulting performance would more closely relate to their proficiency levels, rather than reflect their unfamiliarity with the VWT. They took the VWT 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 correlation matrix between the subscores of each test is given in Table 10. The values across all subscores are at or above $r = .68$. Not surprisingly, the highest correlation coefficients (.96 and .91) exist between subscores (or modules) and the overall scores for the same test. This is true for both VWT and TOEIC.

Table 10. Correlation between for VWT and TOEIC ($N = 55$)

	TOEIC Reading	TOEIC Listening	TOEIC Total
TOEIC Reading	-		
TOEIC Listening	.84	-	
TOEIC Total	.96	.96	-
Versant Writing Test	.70	.68	.72

Though the sample size is small, this matrix shows 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.

The VWT Overall score and TOEIC total score correlated moderately at $r = .72$, as shown in Figure 3, indicating that there is general English ability as a covariance, but that these tests measure different aspects of language performance. The VWT correlated higher with TOEIC Reading ($r = .70$) than with TOEIC Listening ($r = .68$), which is expected because more content is presented through reading than listening in the VWT.

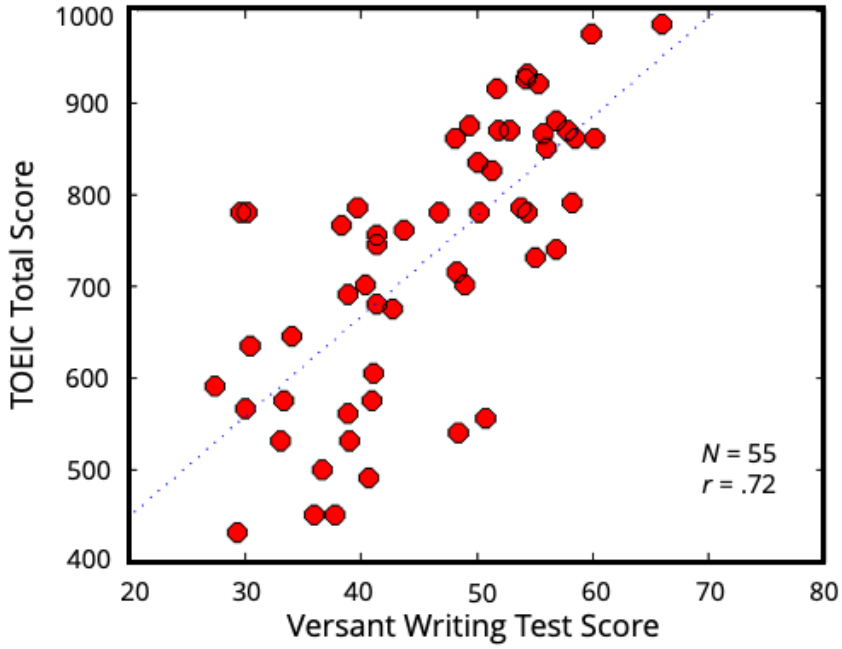

Figure 3. Scatterplot showing the relationship between the VWT and TOEIC (*N* = 55).

7.3.2 VWT and CEFR Level Estimates

In order to identify the correspondence between scores on the VWT 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 VWT scale that maps to the A1 through C2 proficiency levels of the CEFR. A secondary goal of the study was to empirically demonstrate that two item types found on the VWT, Passage Reconstruction and Email Writing, can be reliably evaluated by English language testing experts.

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 a field test version of the VWT 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 1,145 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 = .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 VWT scores 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 11.

Table 11. CEFR score boundaries as logits from a Rasch model

Facet step	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' VWT 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 VWT scores to establish a CEFR Level, 68% of candidates are correctly classified, and 99% of candidates are classified correctly or one step away. Table 12 below provides the final mapping between the two scales.

Table 12. Mapping of CEFR Levels with VWT scores

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

Figure 4 plots the relation between each candidate's VWT 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).

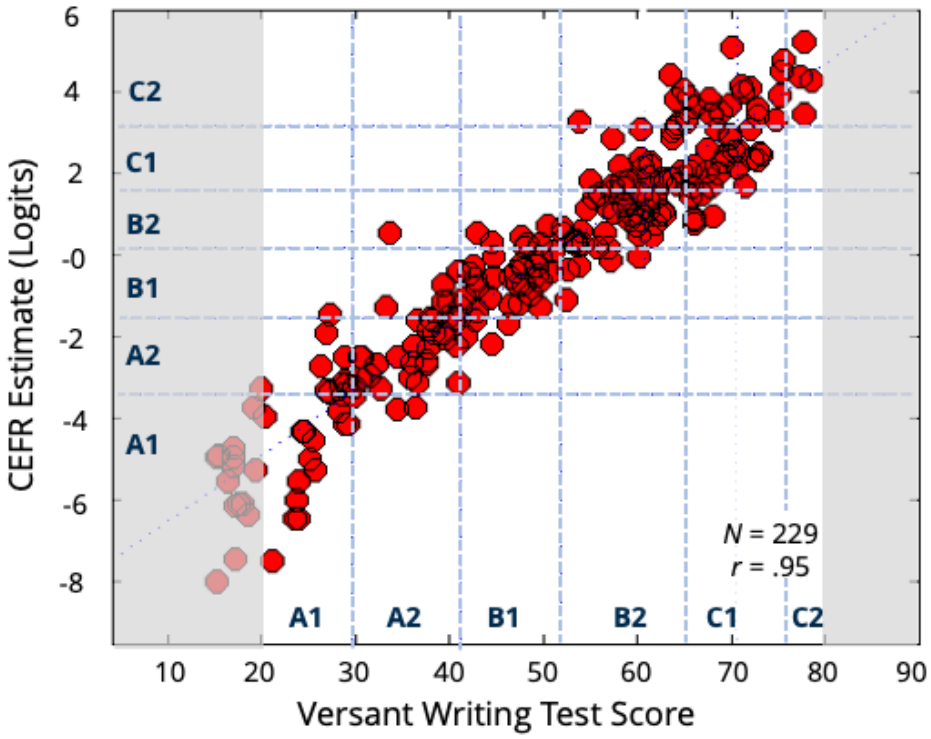

Figure 4. Scatterplot of CEFR ability estimates and VWT scores

The Pearson product-moment correlation coefficients for VWT scores and CEFR estimates is .95, revealing that the VWT instrument 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 VWT Overall 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. Conclusion

This report has provided details of the test development process and validity evidence for the VWT. 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 VWT scores are structurally reliable indications of candidate ability in written English and are suitable for high-stakes decision-making.

9. About the Company

Pearson: Pearson and Ordinate Corporation, the creator of the Versant tests, were combined in January, 2008. The Versant tests are the first to leverage a completely automated method for assessing spoken and written 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 mobile phone 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 VWT 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 investigating new applications for Versant technology. Research results are published in international journals and made available through the Versant website (www.VersantTests.com).

10. References

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

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

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

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

Carver, R. (1991). Using Letter-naming speed to diagnose reading disability. *Remedial and Special Education*, *12*(5), 33-43.

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

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

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

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing*. London: Longman.

Liao, C-W., Qu, Y., & 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, S. (2004). *Assessing Speaking*. Cambridge, UK: Cambridge University Press.

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

Oakeshott-Taylor, J. (1977). Information redundancy, and listening comprehension. In R. Dirven (ed.), *Hörverständnis im Fremdsprachenunterricht. 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.

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

Segalowitz, N., Poulsen, C., & 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.

About Us

We are Pearson English, part of the world's learning company, with expertise in educational courseware and assessment, and a range of teaching and learning services powered by technology.

With 30,000 employees in more than 70 countries, our products are used by millions of professionals, teachers and learners around the world every day. Whether you're a learner seeking swift progress towards new horizons, a teacher who's inspiring achievement in the classroom, an institution looking for measurable improvement, or a professional striving to make data-backed decisions and upskill and reskill their talent for the future, the world of language learning is evolving.

Our mission is to help people make progress in their lives through learning – because we believe that learning opens up opportunities, creating fulfilling careers and better lives.

To try a sample test or get more information, visit us online at:

www.VersantTests.com

Version 0822