WHITE PAPER

The GlobalEnglish Standard Test for English Professionals Plus (STEP+)

Development and Validation

Lindsay Oishi, Ph.D.

Associate Product Owner – Progress and Measurement
GlobalEnglish Corporation

Silvie Liao, Ph.D.

ESL/EFL Learning Specialist
GlobalEnglish Corporation

April 2013

The purpose of this paper is to describe the GlobalEnglish Standard Test for English Professionals Plus (STEP+), an online, computer adaptive test used by global professionals and international organizations to assess Business English proficiency.

First, the paper details the design, development and administration of the test, including its four sections (grammar, reading, listening and speaking), the model of communicative competence on which it is based, and how scores are reported and interpreted. Second, the paper presents information on test reliability and validity, including an empirical validation study, inter-rater reliability, test-retest reliability, and Item Response Theory analyses.

Contents

Introduction

GlobalEnglish STEP+ is an online test designed to evaluate an individual's ability to understand and use English in global business contexts. It is intended for adults and takes about an hour to complete. The test can be taken at any time, at any location, and requires a computer with a standard Internet connection, audio output, and a microphone. A combination of computer and human scoring generates reliable, meaningful results that offer a valid snapshot of an individual's proficiency in Business English.

Business English proficiency is how well a person can comprehend and use English for common business purposes, and includes many aspects of general English language ability. GlobaEnglish STEP+ tests a person's capacity to understand spoken and written English on general and business topics, as well as his or her ability to produce effective spoken English. Corporations and other institutions can use test scores to inform employment, training, and advancement decisions, as well as measure the outcomes of English language programs.

Test Description

Design

GlobalEnglish STEP+ is a computer-based adaptive test that selects a total of 66 questions for each individual test-taker out of an item bank that includes hundreds of questions. Individuals answer multiple-choice questions across four sections: Grammar, Listening, Reading and Speaking (the Speaking section also includes other question types, as shown in Table 1). In all sections except for the Speaking section, there are five difficulty levels of questions, and test-takers receive questions in pairs. All test-takers receive the first pair of questions at Level 3 (Intermediate). If they answer both questions correctly, the next pair of questions is at Level 4. If they answer one question correctly, the next pair of questions is at Level 3. If they miss both questions, the next pair of questions is at Level 2. This process of receiving easier or more difficult questions based on previous answers continues until the final Speaking section. The advantage of adaptive testing is that test-takers receive questions that are appropriate for their level.

There are a total of eleven item types, as summarized in the table below. For every item in the Grammar, Listening and Reading sections, and for Question-Response items in the Speaking section, there are four possible answers and one correct answer. The other Speaking items (reading a paragraph out loud, and open-ended questions) are scored on a range of 0 to 5.

Table 1. STEP+ Sections and Item Types

Section	Item Types
Grammar	Error Detection, Fill in the Blank, Question-Response
Listening	Describe Picture, Question-Response, Short Conversation, and Long Conversation
Reading	General Comprehension, Vocabulary, Pronoun Reference
Speaking	Read a Paragraph, Question-Response (multiple choice with spoken answer), Question-Response (open-ended)

GlobalEnglish STEP+ questions are oriented to the context of adults working in global business, and include functional language on standard topics such as meeting with colleagues, business socializing, telephoning, giving presentations and travel. GlobalEnglish STEP+ does not, however, test knowledge of specific industries, such as manufacturing or finance. Further, it is not limited to a particular geographic variant of English such as British or American English. Audio recordings by men and women feature native English speakers from multiple regions, providing a range of common accents and speaking styles.

Scoring for the first three sections of GlobalEnglish STEP+ (Grammar, Listening and Reading) is by computer, while trained experts in English language and teaching evaluate the Speaking section. Scores are available on a secure website within 48 hours. The GlobalEnglish STEP+ report includes a raw score from 0 to 1,581, and an assignment to an English proficiency level from Beginner to Advanced

(Levels 0 to 10). These assignments also correspond to levels of study in the GlobalEnglish Edge online curriculum.

Administration

Administration of GlobalEnglish STEP+ is via an Internet-capable computer and generally takes about an hour. Test-takers are allowed a maximum of 52 minutes; however, it is not required to use all the time allotted. There are also untimed parts of the test, at the beginning and between each section, during which the test-taker can read instructions, calibrate computer settings, and prepare for the next section.

Each section is timed separately and test-takers cannot use leftover time in one section in any other section. It is also not possible to go back and change any answer or review any previous part of the test. While it is recommended that people take the test in a single session, it is possible to take the test in multiple sittings, either through pausing the test or by exiting and returning later. This does not affect total time allowed, and the test resumes at the last point saved.

Number of Items

The GlobalEnglish STEP+ computerized assessment system presents a total of 66 items in four separate sections to each test-taker. Items are drawn from a large item pool following an adaptive algorithm.

Table 2. STEP+ Sections, Items and Time Allowed

Section	Items Presented	Time Allowed (minutes)
Grammar	26	15
Listening	16	15
Reading	16	15
Speaking	8	7
Total	66	52

Format

This section provides a description of each test section as well as examples and explanations of the types of questions found in each. In all sections, questions are presented one at a time, with an answer required before the individual can move on to the next question. There are equal numbers of each type of item within a section; however, they can be presented at any time (i.e., items of one type are not grouped together).

Part 1: Grammar

The Grammar section includes two types of items: Error Detection and Fill in the Blank.

In an Error Detection question, the test-taker reads a sentence that has four underlined components. Each underlined component is a potential answer choice, and the task is to select which answer is grammatically incorrect. For Fill in the Blank questions, the test-taker must choose the grammatically correct response out of the four potential answers, which are typically similar and represent common errors.

Example: Error Detection

Instructions:

Read the question. Find the mistake.
Click A, B, C or D to choose the answer.

After he received his second promotion, John was the most happy employee in the company.

(A) After
(B) received
(C) most happy
(D) in

Example: Fill in the Blank

Instructions:
Read the question. Click A, B, C or D to choose the answer that best completes the sentence.

A: *Where was Elena this morning? She was supposed to meet us in the conference room at 10:00AM.*

B: *She ___________ an important call regarding the Heyns account.*

(A) must take
(B)had to take
(C) would take
(D) should take

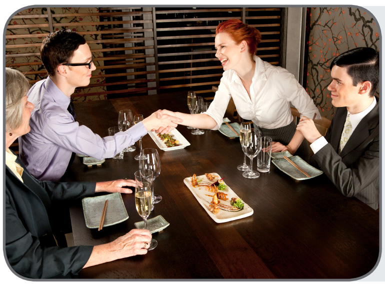
Part 2: Listening

There are four item types in the Listening section: Picture, Question-Response, Short Conversation, and Long Conversation.

In the Picture questions, the test-taker sees a photograph of a common social or business situation, and listens to four answer choices being read aloud. The answer choices are not displayed as text, and the test-taker can only listen to them twice.

Example: Picture

Instructions:
Look at the picture. Click PLAY and listen (You can listen only twice). Click on A, B, C or D to select the best answer.



PLAY

○ (A)
○ (B)
● (C)
○ (D)

Answer choices (not displayed):

(A) The man is drinking coffee. (B) The women are shaking hands. ***(C) The man and woman are meeting at a restaurant.*** (D) The men are drinking wine.

In Question-Response items, test-takers listen to a spoken question and select the written answer that is the most appropriate response. The question is not displayed as text and the test-taker can only listen to it twice. There is a photo along with the question to provide context, but it is not required to answer the question correctly.

Example: Question-Response

Instructions:
Listen to the question and select the most appropriate response to that question.

PLAY

Question (not displayed): "How often do you play golf?"

(A) Yes, let's play next week.
(B) I play about twice a week.
(C) I'll probably golf today.
(D) I play golf with my dad.

In Short and Long Conversation items, test-takers listen to a conversation between two people and then answer a written question about that conversation. The answer choices are also presented as text. Short conversations typically involve one sentence given by each person in the dialogue. Long conversations typically involve more than one sentence by each person in the dialogue. The conversations are not presented as text and test-takers can only listen to each item twice.

Example: Short Conversation

Instructions:
Listen to the conversation between the two people. Then, answer the question about the conversation you heard.

PLAY

What does the man offer to do?

(A) Ask the woman to call Ms. Jones immediately
(B) Call Ms. Jones again the next day
(C) Request that Ms. Jones return the call later
(D) Invite the woman to call Ms. Jones in the morning

Part 3: Reading

The Reading section is comprised of General Comprehension, Vocabulary, and Pronoun Reference items in approximately equal numbers. General Comprehension questions test a reader's ability to recognize explicit information, make reasonable inferences from a text, and use this knowledge to select a written answer with appropriate semantic content in response to a written question. Vocabulary items require the test-taker to identify a synonymous word or phrase for a selected word from the reading passage, as it is used in that context. Finally, Pronoun Reference items ask the test-taker to identify the noun that a pronoun such as "it", "he" or "they" refers to. In easier questions, this noun is clearly stated in the passage; in harder questions, the noun may have to be inferred from the text. As the reading passages are short but complete texts, typically a few hundred words long, there are two questions for each passage.

Example: Reading Passage

Annual Sales Report

Dear Dymo Employees,

We had an excellent year at Dymo Industries. Total sales were US$550 million. This is an impressive growth of 10% compared to last year.

There are a few important reasons for the large increase in sales. All the international divisions performed very well. The new sales training helped many of us promote our new line of toys. In addition, the creation of new products opened more markets for us.

We continued last year to provide our customers with improved products. We all know that product enhancement can lead to larger market shares. Therefore, we hope to continue increasing our sales by creating better products in the coming year.

Congratulations on your great teamwork!

Jon Ngyuen

Example: General Comprehension

Instructions:
Read the document below. Then read the question. Click A, B, C or D to select the best answer.

Which of the following is NOT a reason for Dymo Industries' increased sales?

(A) A new sales team
(B) New markets
(C) Sales training
(D) Performance in international divisions

Example: Vocabulary

Instructions:
Look at the word that is highlighted in yellow in the reading passage. Choose the word or words that have the same meaning as the highlighted word.

Look at the word "divisions" in the article. Which words or phrases mean the same thing as "divisions" in this article?

(A) Disagreements within the company
(B) Groups or teams within an organization
(C) Sales methods
(D) Individual employees

Example: Pronoun Reference

Instructions:
Instructions: Look at the pronoun that is highlighted in yellow in the reading passage. Choose the word or words that have the same meaning as the highlighted pronoun. The word or words will come before the highlighted pronoun. Some may even be in the previous sentence.

Look at the word "this" in the article. What does the word "this" refer to in the previous sentence?

(A) Total sales
(B) Dymo Industries
(C) Last year
(D) Division performance

Part 4: Speaking

In the speaking section, examinees read a written paragraph out loud, answer five short questions (where responses are provided in text), and respond to two open-ended questions. Test-takers are instructed to speak naturally, using correct pronunciation, intonation, stress and rhythm. Test-takers can only record their answers once, and while recording, they cannot pause the timer. There is an exception to this rule if the audio recorded is too poor to produce a score-able response; in that case, the test-taker is allowed to re-record his or her response. If test-takers produce two recordings without intelligible audio, the test takes them back to the microphone calibration tool.

In Part A, test-takers read a short paragraph aloud. Once they begin recording, they have 60 seconds total to finish reading the paragraph aloud; however, they may stop the recording at any point if they finish before 60 seconds are used. Part A has one question and is designed to test pronunciation, including the ability to say all of the sounds of English, stress the correct syllables, pause appropriately, and use the correct intonation. This allows scorers to evaluate the candidate's ability to produce intelligible and comprehensible speech. Intelligibility refers to the production of sound patterns that are recognizable as English. Comprehensibility refers to the ability to speak so that people can understand the meaning. Test-takers read paragraph aloud that has all of the phonemes of English and a variety of sentence types.

Example: Read Paragraph

Instructions:
Click RECORD and read the paragraph aloud. Click STOP when you are finished.

● RECORD

■ STOP

Our CEO announced a decision today at the monthly company meeting. The executives think that we should expand our product line. We just need to add a few luxury items to some electronic goods. That should help us increase our market share in Europe and Asia. Future profits depend on our dealing with this issue now. We hope to increase sales by 10% in the next year.

In Part B, test-takers must listen to audio cues, choose a written response that is appropriate in the context, and speak it. They are evaluated on their ability to choose the correct answer and to say it clearly. Part B has five questions and is designed to test communicative competence. Communicative competence refers to a speaker's ability to apply linguistic, discourse, and cultural knowledge to produce appropriate language in a particular context.

Example: Select & Speak Response

Instructions:
Click PLAY and listen. You can only listen twice. Click RECORD and then read the correct response aloud. Click STOP when you are finished.

PLAY

● RECORD

■ STOP

Audio Cue (not displayed): "We don't have time to discuss everything on the agenda at today's meeting."

(A) The meeting started an hour ago.

(B) What's the next point on the agenda?

(C) That's okay, we can continue tomorrow.

(D) Well, what in particular concerns you?

Part C has two open-ended questions and is designed to test overall linguistic performance. Open-ended questions allow evaluation of how well test-takers construct a response to show evidence of their acquired knowledge. The first question requires only a beginner to low intermediate level of English to answer adequately, while the second question requires greater topic development and linguistic complexity, and demonstrates high intermediate to advanced ability when answered well. Test-takers are given 75 seconds for the low-level response and 105 seconds for the high-level response, including preparation time.

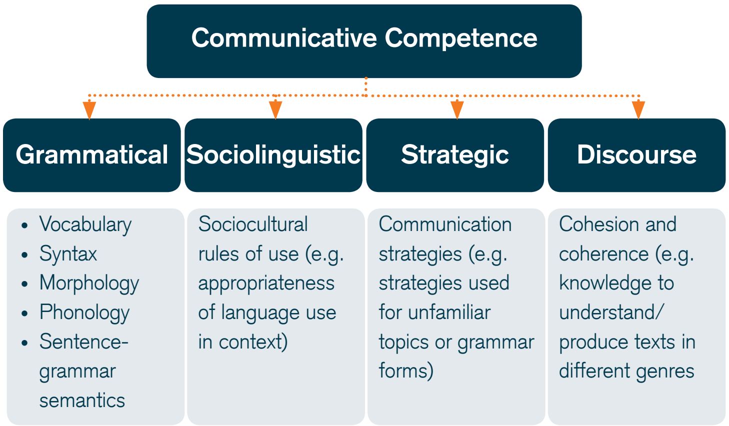
Table 3. Examples of Open-ended Questions

Level	Question
Low	What is your favorite hobby or activity? Why?
High	How does the Internet affect your life? What are the advantages and disadvantages of the Internet?

Test Construct

Canale and Swain's (1980) model of communicative competence is one of the most influential frameworks in language testing, and argues for the need to assess grammatical, sociolinguistic, strategic and discourse competence. In this model, simple declarative knowledge of the language (e.g., being able to state the past tense of a verb) is not sufficient evidence of ability to actually use the language for real communicative purposes. Figure 1 presents the theoretical framework of this model. Tests of communicative competence not only reflect the need for language to be used interactively with other speakers of that language, but also incorporate authentic contexts, situations, and topics. In this way, tests based on this model aim to evaluate and individual's overall ability to communicate successfully in real-life situations.

Figure 1. STEP+ Test Construct


Communicative Competence
Grammatical
Sociolinguistic
Strategic
Discourse
• Vocabulary
• Syntax
• Morphology
• Phonology
• Sentence-grammar semantics
Sociocultural rules of use (e.g. appropriateness of language use in context)
Communication strategies (e.g. strategies used for unfamiliar topics or grammar forms)
Cohesion and coherence (e.g. knowledge to understand/ produce texts in different genres

As evident in the figure above, communicative competence is a complex construct that involves more than one language skill (such as reading or speaking per se). To provide a strong evaluation of a candidate's actual communicative ability, a test must therefore address multiple language skills. Figure 2 shows the language skills assessed in GlobalEnglish STEP+, and Figure 3 illustrates how each section within the test addresses multiple facets of communicative competence. As a comprehensive test, GlobalEnglish STEP+ is designed to measure more than English language knowledge or general linguistic skill. Instead, it specifically targets the ability to effectively use English in global business contexts.

Figure 2. Language Skills Assessed by STEP+

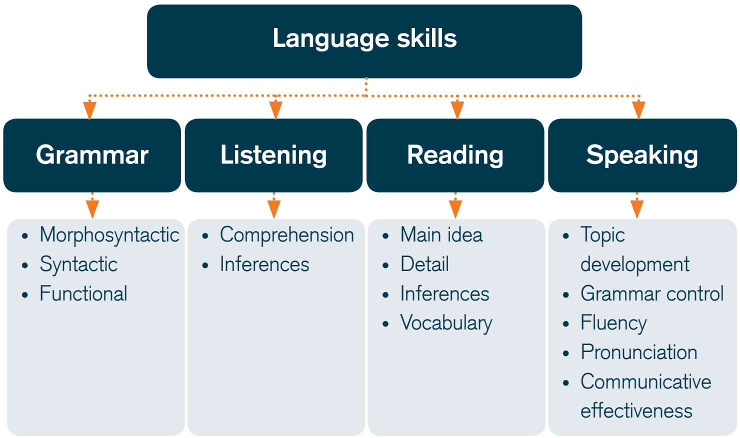
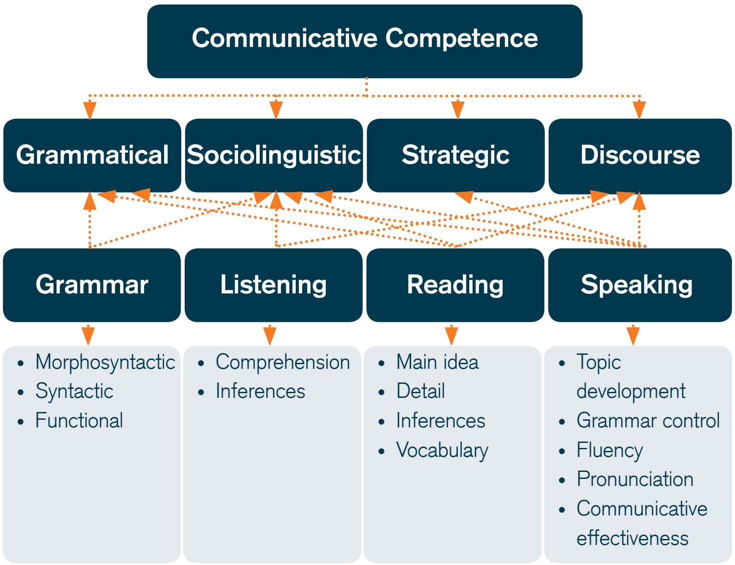


Figure 3. Communicative Competence and Language Skills Tested in STEP+



Content Design and Development

The GlobalEnglish curriculum and STEP+™ were developed by a team of experts with extensive training and experience teaching English as a foreign language, creating curricula, and writing coursework. GlobalEnglish Corporation's world-class Academic Advisory Board, including four former presidents of International TESOL, aided this team. Dr. David Nunan, the senior academic advisor to GlobalEnglish, was director and chair of applied linguistics at the University of Hong Kong, and the president of International TESOL in 1999-2000. He has published over 100 books and articles in the areas of curriculum and materials development, classroom- based research, and discourse analysis. For more information, visit www.globalenglish.com/whoweare/aboutus/academic_advisors.

Item Development and Vocabulary Selection

As GlobalEnglish STEP+ was designed to align with GlobalEnglish levels, test items were created with reference to content and learning objectives from the levels, which are available by request or through the GlobalEnglish website. The vocabulary used in GlobalEnglish STEP+ was also based on the GlobalEnglish courses, and test-writers were given a comprehensive list of 3,150 vocabulary words, as well as the syllabus for each level.

For the grammar section, test writers were given the test framework, which listed grammatical features across different English proficiency levels, and were advised to provide a balance across the three types of grammar features: morpho-syntactic (e.g. articles "a/an/the"), syntactic (e.g. use of a preposition), and functional (e.g. tense marking in conditionals). For the listening section, test writers were given a list of common categories of expressions in business communication (e.g. asking for and giving opinions). For the reading section, test writers worked from a list of genres common in the workplace, such as memos, forms, charts and announcements.

Test items in the speaking section underwent a different procedure to create the three distinct sub-sections. In Part A (reading paragraphs), test writers were advised to write paragraphs that met the following criteria:

- Length of 65 to 75 words.
- Score between 75 and 85 on the Flesch Reading Ease scale.
- Paragraph subjects related to business.
- Contain at least two instances of every phoneme in English.
- A variety of consonant clusters (e.g. the "lpf" in the word helpful).
- Contain words with the phonemes /s/ and /z/ at the end (e.g. boss, was).
- Some variation in sentence intonation (e.g., questions and statements).

Items in Part B and Part C were adapted from questions in the GlobalEnglish curriculum. After those speaking items were drafted, 30 non-native speakers of English were recruited to take a pilot test. Based on their responses, items that seemed inappropriate, confusing or dependent on factors such as test-takers' working experiences, regions or familiarity with certain English dialects were either eliminated or revised.

Overall, the GlobalEnglish STEP+ item bank was designed so that any two randomly selected test forms would have similar test content and difficulty. The item bank has also undergone two thorough reviews. Dr. Kathleen Bailey[1], a recognized expert in English language testing, analyzed the item pool, checking item difficulty, and examining each item for its correspondence to the test construct. Dr. Jean Turner[2] of the Monterey Institute, a recognized expert in English language testing, then conducted a statistical analysis of each test item's discrimination (capacity to separate low-level examinees from high-level

[1] Dr. Kathleen Bailey is a professor of applied linguistics at the Monterey Institute of International Studies in Monterey, California. Director of the TESOL M.A. program there for six years, she was also the Director of the Intensive English as a Second Language Program for three years. Dr. Bailey was the 1998-1999 president of TESOL. She has worked with language teachers in many countries and is the author or coauthor of numerous scholarly articles and books. Her two most recent books are Pursuing Professional Development: The Self as Source (co-authored with Andy Curtis and David Nunan, Heinle & Heinle, 2001) and Practical English Language Teaching: Speaking (McGraw-Hill, 2004). Dr. Bailey earned her Ph.D. in applied linguistics from the University of California, Los Angeles.

[2] Dr. Jean Turner is a professor of applied linguistics at the Monterey Institute of International Studies. She is a recognized expert in the field of English language testing. She received her Ph.D. from the Graduate School of Language and Educational Linguistics, Monterey Institute of International Studies.

examinees, based on their overall pattern of responses). Items with appropriate item discrimination were retained in the GlobalEnglish STEP+ item bank, while others were revised under Dr. Bailey's guidance. This effort reduced measurement error, increased reliability, and enhanced construct validity.

Item Recording

Sixteen professional voice talents (7 males and 9 females) were hired to record listening items and speaking prompts. All voice actors were native speakers with a range of common American and British English accents. Recordings were made in a professional recording studio in South San Francisco, California. An ESL expert was on site to ensure that the speed and pronunciation were clear, natural and appropriate to the items at different levels of difficulty.

Score Reporting

Scoring for the first three sections of GlobalEnglish STEP+ (Grammar, Listening and Reading) is computerized, and when the audio files from the Speaking section have been scored (within 48 hours and often sooner), the test-taker receives an email with a link to a secure results web page.

The Score Report includes:

- A raw score from 0 to 1,581.
- The test-taker's recommended English level at which to begin studying (i.e., Level 3: High Beginner), and a description of that level.
- Performance (raw score) in each of the four sections: Grammar, Listening, Reading, and Speaking.
- A description of strengths and weaknesses by section.

Scores and Weights

GlobalEnglish STEP+ has a score range of 0 to 1,581, calculated by summing the scores for each of the four sections. The maximum possible scores vary by section: 403 in Grammar, 392 in Listening, 395 in Reading and 391 in Speaking.

The raw score for each section is the sum of points awarded for correct answers. Easier questions have lower point values, and harder questions have higher point values, with a range of 2.5 (easiest) to 27.5 (hardest) points awarded for a single correct answer. Total score for the grammar section is divided by 1.625 in order to provide an equal weighting of grammar with listening and reading sections (the grammar section has 26 questions, while the listening and reading sections each have 16). It is possible for a user to be awarded more points than the maximum section score. In this case, the section score is set to the maximum.

The Speaking section is evaluated by trained, expert human scorers with extensive experience in teaching and assessing English as a foreign language. There are three parts to the Speaking section, with a different question type in each.

Scoring for Part A of the Speaking section (reading a paragraph aloud) is holistic and depends on two key factors: intelligibility and comprehensibility. Scorers are instructed to think about the extent to which pronunciation and fluency problems reduce understanding of what the test-taker said, and to gauge how much any accent interferes with basic intelligibility. Each question is given a score of 0.0 to 5.0, with half-points allowed.

Part B of the Speaking section contains multiple choice questions in increasing order of difficulty, and all questions are scored as correct or incorrect.

Scoring for Part C of the Speaking section (two open-ended questions) is holistic, but the rubric includes consideration for grammar and topical development (how completely and well the response answers the question), as well as pronunciation and fluency. Each of these four sub-scores is given a score of 0 to 5, with no half-points allowed.

Score Use

GlobalEnglish STEP+ is a diagnostic tool that can be used in multiple ways. First, it is a valid indicator of the most appropriate level for an individual to begin studying

Business English. Second, organizations can use the test to estimate the overall Business English proficiency of a potential or current employee, and use this information to inform hiring, training and advancement decisions. Third, GlobalEnglish STEP+ can be used to ascertain progress in Business English after a significant and sustained period of study.

Scores on GlobalEnglish STEP+ are valid and reliable indicators of Business English proficiency when the test is taken in compliance with recommended conditions. We strongly suggest that test-takers complete the test to the best of their ability, with no references, aides or outside help, in a single session. An organization that wishes to enforce these recommendations or require evidence of an individual's identity or independent performance should set up a monitored, formal administration procedure or use a respected third party administrator.

Score Interpretation

GlobalEnglish STEP+ raw scores are reported along with a GlobalEnglish Business English Proficiency Level from 0 (Low Beginner) to 10 (Advanced). Table 4 shows how GlobalEnglish levels align with GlobalEnglish STEP+ scores.

These scores can also be interpreted using the Common European Framework of Reference (CEFR, Council of Europe, 2001), a widely used standard that describes six broad levels of proficiency in any language. The CEFR was designed to provide common reference points and terminology for learners, teachers, companies and other institutions to describe achievement across diverse educational contexts. Table 4 also shows how GlobalEnglish STEP+ scores relate to CEFR levels and gives an example descriptive statement for level. These relationships are intended as general guidelines, not specific claims about the abilities of GlobalEnglish STEP+ test-takers.

Table 4. STEP+ Scores and the Common European Framework of Reference

GlobalEnglish STEP+ Score	GlobalEnglish Level	CEFR Level	CEFR Description
0 - 120	0	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.
121 - 240	1		
241 - 347	2	A2	Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters.
348 - 472	3		
473 - 620	4		
621 - 768	5	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.
769 - 916	6		
917 - 1064	7		
1065 - 1212	8	B2	Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.
1213 - 1399	9		
1400 - 1581	10	C1	Can use language flexibly and effectively for social, academic and professional purposes.
N/A	N/A	C2	Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.

** For complete global descriptors for CEFR levels, please see the Common European Framework of Reference for Languages, available in English at http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf*

Validation

As GlobalEnglish STEP+ is continuously monitored, evaluated, and revised, validation of the test is an ongoing process. The original test development team was comprised of English language teaching and assessment experts who drew on extensive knowledge of the Common European Framework and other standards to create items with appropriate content and difficulty. In 2011, a respected external firm was hired to perform a review of item performance and the test overall; as a result, underperforming items were removed or revised in early 2012. In late 2012, the speaking section of GlobalEnglish STEP+ was revised significantly in order to offer a more complete evaluation of speaking ability, with more challenging questions and a more comprehensive scoring process. To understand the consequences of the changes to the test in early and late 2012, data from that year was used to analyze test reliability and validity.

Study Design

Data on test reliability and validity was analyzed from three samples:

- For overall test reliability and item analysis using the Item Response Theory (IRT) model, test results were used from a three-week period in November and December 2012 (N = 2,862).
- For test-retest reliability analysis, test results were taken from May 2012 through March 2013 and included those test-takers who took the test twice in one day (N = 384).
- For the study on how test results related to other constructs, participants (N = 62) were employed, adult volunteers who took GlobalEnglish STEP+ in October and November of 2012. Participants completed a short survey with demographic and other questions after they finished the test. There were 16 men and 28 women in the sample (18 did not state gender), with an average age of 33 (range: 20-55 years). Nine native English speakers were included in the validation study for comparison purposes. Participants were from a variety of countries in Europe, South America, and Asia, and had a range of professions.

Reliability

Overall Test Reliability

As GlobalEnglish STEP+ is a computer adaptive test with a large pool of potential questions, it is not practical or useful to compute overall test reliability metrics used in classical test theory (such as Cronbach's alpha). A modern alternative approach is Item Response Theory (IRT), which considers the contribution of each question to the overall test's capacity to give valid and reliable scores. IRT is "a diverse family of [psychometric] models designed to represent the relation between an individual's item response and underlying latent trait," (Fraley, Waller and Brennan, 2000) that is widely used in educational testing, and useful for modeling communicative competence in English (a latent trait).

As IRT analysis is designed for tests that measure a single latent variable, a factor analysis of principal components (PCA) was first performed on the four GlobalEnglish STEP+ subscores (Grammar, Listening, Reading, Speaking) to ascertain unidimensionality across the test. As shown in the screeplot presented in Figure 4, the analysis indicated a single-factor solution, with the first factor explaining 63% of the variance in subscores. A factor map of the variables showed that while all subscores loaded positively on the first, primary factor, the Speaking subscore also loaded positively on a second factor (Figure 5).

Figure 4.
Screeplot of Principal Components using STEP+ Section Scores

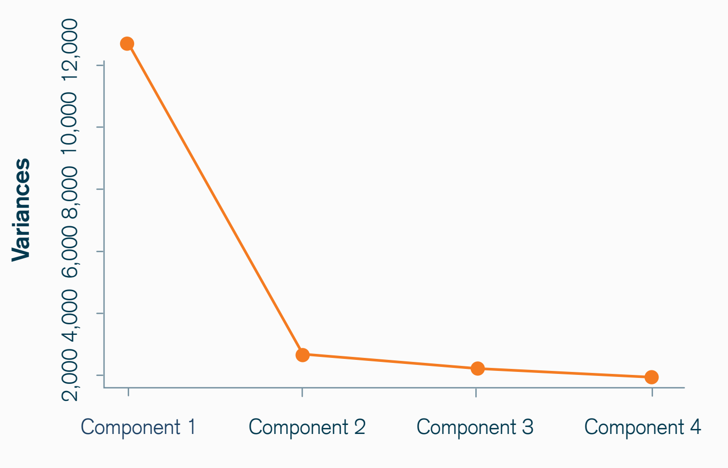
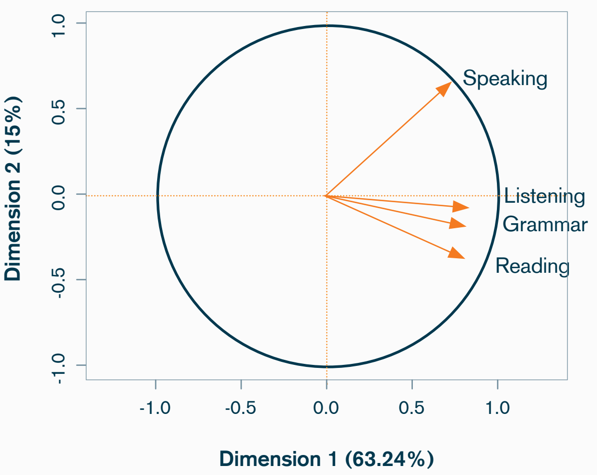


Figure 5.
Variables Factor Map (Principal Components Analysis)



For the validation study, we used the Rasch model, which is commonly used in IRT for educational tests (Rasch, 1960/1980). To fit the model, we used the maximum log-likelihood method as specified in the ltm package of the R statistical analysis software (Rizopoulos, 2012). As the IRT analysis with the full data set (767 items, 2,862 observations) produced unstable solutions due to a large number of missing values, the dataset was reduced to the most frequent items (456 items, 60% of total) and the most complete observations (2,053 observations, 72% of total). Although a complete description of the IRT results is beyond the scope of this white paper, using the IRT model created with this dataset, a simulation of 1,000 possible test forms was performed. Across the 1,000 test forms, average test information = 41.29, SEM = .16, and reliability = .97. The model also indicated that while the test was informative (reliable) at all skill levels, it was most informative for test-takers at the intermediate to high intermediate proficiency levels.

Test-Retest Reliability

Another method for ascertaining test reliability is to determine whether a person who takes the test more than once will obtain a similar score (absent intervening study or other factors). To perform this analysis, we examined a sample of test-takers who voluntarily and spontaneously chose to take a test composed of Reading, Listening and Grammar items from the STEP+ pool (this test is available as the Business English Placement Test to GlobalEnglish Edge subscribers). Among those who took both tests in the same day (N = 384), correlation between their two test scores was positive and significant ($r = .82$, $p < .001$).

Inter-rater reliability

When the GlobalEnglish STEP+ Speaking section was revised in September 2012, the team of expert scorers was extensively re-trained to address the addition of open-ended questions. As part of the training, scorers listened to sample responses, which represented a range of English language proficiencies, and discussed scoring issues until they reached agreement. An initial manual was composed to guide scoring for the validation period.

During validation, scorers were randomly assigned to two groups, and each group scored the same 29 tests (which were also randomly assigned within each group), with a total of 394 Speaking items after unusable audio files were removed. After both groups had scored the tests, items with significant score differences were identified and reviewed. The GlobalEnglish STEP+ team discussed scoring issues with all scorers, and the scorer manual was revised accordingly. Finally, the 25 items with the greatest discrepancies were re-scored by their original scorers following the revised guidelines, and inter-rater reliability was computed using these final scores. Inter-rater reliability across the entire Speaking section was .84.

Validity

Dimensionality

One indication of validity is a test's ability to reveal multiple dimensions of a particular construct, such as language, which is theorized to contain sub-skills or other secondary components. While GlobalEnglish STEP+ tests Business English language proficiency generally, it also should allow for test-takers to have different levels of achievement in four linguistic sub-skills: grammar, listening, reading and speaking.

In Table 5, below, the correlations between sub-scores on GlobalEnglish STEP+ show that while each sub-score was positively and significantly related to the other sub-scores, these relationships were of moderate strength (Pearson's r-values of .42 to .57). Higher r-values would suggest that GlobalEnglish STEP+ sub-sections were providing redundant information on highly similar constructs, while lower r-values would indicate too much disparity in sub-skills that are theoretically linked. It is also worth noting that receptive skills (grammar, listening and reading) are more strongly related to each other than they are to productive skills (speaking), as would be expected in standard models of language learning. Overall, the moderate correlations demonstrated here support the validity of GlobalEnglish STEP+ as a Business English assessment with the ability to discriminate between different secondary language skills.

Table 5. Correlations for STEP+ Section and Total Scores (N = 2,862)

	Listening	Reading	Speaking	Total
Grammar	.57	.57	.48	.84
Listening		.53	.48	.81
Reading			.42	.82
Speaking				.70

All p's < .01.

Native and Non-native Performance

Although the number of native speakers in the validation sample was small (n = 9), their performance on GlobalEnglish STEP+ was high, as expected. The average total GlobalEnglish STEP+ score for native speakers was 1,414 (out of 1,581), with a standard deviation of 68 points. For non-native speakers (n = 44), however, the average was 1,012, with a much greater standard deviation of 300 points. These results, while preliminary, are evidence for the construct validity of GlobalEnglish STEP+ as a reflection of English language ability. As we continue to work on validation, more native speakers will be added to this sample.

Relationship to Self-Rating

As a further test of construct validity, we examined the relationship between self-ratings and performance on GlobalEnglish STEP+. While certainly imperfect, self-ratings are a reasonable way to approximate the correspondence of a test with what it measures, in domains where individuals can understand their own proficiency and report it with some degree of accuracy. Table 6 shows that among participants in the validation sample who had both self-ratings and complete GlobalEnglish STEP+ scores (n = 44), there were moderate, positive correlations between the two variables, both overall and for different language skills (corresponding to sections within the test).

Table 6. Correlation between STEP+ Section Scores and Self-Ratings

STEP+ Section	Correlation of Score and Self-Rating
Overall	.53**
Grammar	.38*
Listening	.42*
Reading	.45*
Speaking	.65*

** p < .01; ** p < .001*

Relationship to Previous Study

The relationship of GlobalEnglish STEP+ scores with previous English language study was also examined as evidence of construct validity. As language proficiency is a learned skill, the duration and quality of previous study should have an impact on test outcomes. We asked validation study participants to report how much time they spent studying English in various contexts, including primary, secondary, and post-secondary education, as well as independently through private instruction and self-study. Table 7 shows that the self-reported average amount of time (in years) spent studying English across all categories was positively correlated with overall performance on GlobalEnglish STEP+, Spearman's r (36) = .46, *p* < .001. Time spent studying English during post-secondary education, graduate school, and private instruction were all significantly related to GlobalEnglish STEP+ scores, while English language study during primary and secondary school, or through self-study (including via the Internet) were not significantly related to GlobalEnglish STEP+ scores.

Table 7. Previous Language Study Associated with STEP+ Scores

	Educational Context (Years of English Study)	N	Correlation with STEP+ Score
Significantly Related to STEP+			
	Private Instruction	27	**.48***
	Lifetime Average (All Instruction Types)	36	**.46****
	Post-secondary Education	32	**.41***
	Graduate Education	32	**.39***
Not Significantly Related to STEP+			
	Internet-based Study (not including GlobalEnglish Edge)	28	.31
	Primary Education	33	.28
	Self-study	28	.20
	Secondary Education	33	.15

* *p < .05; ** p < .01*

Conclusion

In today's global economy, Business English is vital to organizational productivity. Companies and institutions need to be able to assess Business English as part of critical human capital decisions such as recruiting the right person for the right job, handling promotions and raises, ensuring personnel skills development, and retaining employees. GlobalEnglish's robust, online and adaptive assessment enables organizations to evaluate and monitor the language skills of their employees and job candidates with the end-goal of increasing organizational communicative and collaborative capacities.

When used as recommended, GlobalEnglish STEP+ is an effective diagnostic of Business English proficiency with significant advantages over alternative assessments. As an online test, it is always available and does not require travel to a test center at a specific date and time. Its breadth of question types allows for a comprehensive picture of Business English capability, yet because of its adaptive design, the test takes an hour or less. Test-takers receive detailed feedback on their performance, and Administrators can easily access the score information and speech samples that they need. GlobalEnglish STEP+ is the best solution for companies wanting fast, accurate, easy-to-implement evaluations of Business English competence.

About the Authors

Dr. Lindsay Oishi is the Associate Product Owner for Progress and Measurement at GlobalEnglish Corporation. She has a BA from Georgetown University, an MA from Oxford University, and a PhD in Educational Psychology from Stanford University. Dr. Oishi has worked on research, assessment and evaluation in both academic settings, at the Hasso Plattner Institute of Design and the Stanford University Online High School, and in business settings, at Hewlett-Packard and Adobe Systems. She has written about online education and language learning in publications for various organizations, including the International Association for K-12 Online Learning, the California Foreign Language Project, and the American Educational Research Association.

Dr. Silvie Liao is the ESL/EFL Learning Specialist at GlobalEnglish Corporation. She is an expert linguist, with an MA in TESOL from Teachers College, Columbia University, and a PhD in Linguistics from the University of California, Davis. Throughout her academic career, Dr. Liao has published articles on second language acquisition, sociolinguistics, language variation, and related topics. She has over ten years of teaching experience with students of all ages from diverse language and cultural backgrounds.

About GlobalEnglish

GlobalEnglish offers an innovative, cloud-based suite of Business English solutions specifically designed to produce immediate productivity and performance gains enterprise-wide. In blending the latest technology innovations with research on how adults effectively acquire language, GlobalEnglish provides a comprehensive solution: formal and informal Business English learning, instant on-the-job support for business tasks in English, enterprise collaboration, mobile productivity, adaptive Business English assessments, and the ability to measure usage and proficiency improvements across the company. GlobalEnglish experts located throughout the world help companies maximize the value of their investment through custom analysis and recommendations, coordinated program deployment, and ongoing support in 15 languages. Headquartered in Brisbane, California, GlobalEnglish has partnered with more than 500 of the world's leading corporations and enterprises, including BNP Paribas, Capgemini, Deloitte, GlaxoSmithKline, Hilton, John Deere, Procter & Gamble and Unisys. GlobalEnglish is owned by Pearson, the world's leading learning company.

References

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards and R.W. Smith (Eds.), *Language and Communication*. New York: Longman.

Canale, M., Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

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

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual.* Strasburg: Language Policy Division.

Hymes, D.H. (1966). Two types of linguistic relativity. In Bright, W. (Ed.), *Sociolinguistics*. The Hague: Mouton.

Fraley, R.C., Waller, N.G., and Brennan, K.A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*, 350–365.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Rizopoulos, D. (2012). *Latent Trait Models under IRT.* Retrieved from http://rwiki.sciviews.org/doku.php?id=packages:cran:ltm.