ETS®

TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

Background and Goals of the TOEIC® Listening and Reading Test Redesign Project

Mary Schedl

January 2010

The TOEIC® Listening and Reading test is an assessment of proficiency in English language as it is used in the global workplace, where English is the means of communication among both native and non-native speakers. Since 1979, the TOEIC test scores have informed decisions regarding recruitment, job placement, promotion, and training, and the test has been widely recognized as a worldwide standard in the assessment of international English use.

In January 2003, a team of content and statistical analysis specialists was formed to consider a redesign of the TOEIC Listening and Reading test as it then existed. The test redesign coincided with an effort to investigate the possibility of making high quality TOEIC Speaking and Writing tests available to test score users as additional test components. Information about the development and design of the constructed response modules is available separately. This document focuses on the redesign of the TOEIC Listening and Reading test.

The goals of the redesign project were:

1. bring the test into alignment with current theories of language proficiency,
2. identify the major variables contributing to the difficulty of language tasks on the revised test using evidence-centered design (ECD) methodology to, and
3. provide more proficiency information that is meaningful to test takers and score users.

The TOEIC redesign team was able to utilize and build on the theoretical and research base for English language learning and testing that had been developed over the course of a decade by the TOEFL iBT™ project. The history and findings of that project have been documented in detail and published in book form. The theoretical underpinnings of language proficiency are documented in the new TOEFL listening and reading frameworks (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, & Schedl, 2000) and numerous monographs.

Aligning the Test With Current Language Theory

Current language theory recognizes the complexity of authentic language contexts; in these contexts, it is often necessary for the learner to use multiple abilities and strategies in order to comprehend and connect information that is heard and read. Communication in real-world situations usually requires the simultaneous engagement of lexical, grammatical, phonetic and pragmatic language abilities. While the Classic TOEIC Listening and Reading test included brief spoken and written text samples, it also included a number of very short, single sentence contexts that focused on discrete language abilities. Table 1 illustrates the components of the TOEIC Listening and Reading test sections prior to the redesign.

TABLE 1

The TOEIC Listening and Reading Test Before Redesign

Test section	Item type	Number of items
Listening	Photographs	20
	Question-response	30
	Short conversations	30
	Short Talks (sets of 2-4 items)	20
Reading	Incomplete sentences	40
	Error recognition	20
	Reading passages (sets of 2-4 items)	40

Proposed revisions to this format included:

1. lengthening some of the listening and reading stimuli
2. varying the accents in the recorded stimuli in the listening test
3. eliminating the error recognition item type from the reading section
4. reducing the number of photograph items in the listening section
5. replacing some of the error recognition items in the reading section with passage-based sentence completions
6. including some sets of items in the reading section that were based on two related passages

Lengthening the Listening and Reading Stimuli

The redesign team found that most of the item types represented in the TOEIC test did require examinees to understand language in context, whether in a single spoken or written sentence or a longer text, and to use multiple language abilities and strategies. However, because the team recognized how important it is in many real-world situations to process longer spoken and written texts, the team considered lengthening some of the listening and reading stimuli. This would create more opportunities for examinees to demonstrate their ability to understand language not only at the sentence level but also within a larger context. Consequently, one of the research questions posed by the redesign team was whether longer listening and reading texts could be used to broaden the measurement of examinee abilities and increase the authenticity of the TOEIC Listening and Reading test. The proposed redesign would increase the number of items associated with each stimulus and reduce the total number of stimuli. An additional advantage of using longer stimuli with more questions per stimulus was that the number of different topics examinees needed to process would be reduced.

Varying Accents

Another revision to the stimuli (the dialogues, conversation, and talks) of the listening section was proposed as a means to improve authenticity. The Classic TOEIC Listening and Reading test included only standard American accents. The design team proposed varying the accents used in the listening section because the TOEIC test is a test of international English, and examinees are likely to need to communicate with a variety of English speakers. With different English accents represented in the test, the redesigned test would reflect the varieties of standard English that listeners would likely be exposed to both in learning English and in hearing it spoken in the international workplace.

Eliminating Error Recognition Items

Of the item types in the Cassic TOEIC Listening and Reading test, the error recognition items were determined to be the least appropriate from the point of view of current language proficiency theories. These items required examinees to identify errors in discrete sentences. Rather than measuring the ability of examinees to understand authentic communication, the items measured the ability to recognize lexical and grammatical errors in written language. This item type was tentatively marked for exclusion from the Redesigned TOEIC Listening and Reading test.

Three other revisions were proposed to test item content to meet the first project goal of modernizing the test by bringing it into alignment with current theories of language proficiency.

Reducing the Number of Photograph Items

The redesign team decided to reduce the number of discrete single-sentence photograph items in the listening section from 20 to 10 in prototype forms, pilot tests, and finally in the field study. The photograph items were typically the easiest items on the test, and they were also, therefore, less discriminating for the majority of examinees than other item types. The decision to keep half of them was made for several reasons: some very easy items were appropriate to discriminate lower-level examinees; the photograph items add visual interest to the test; and the items are useful for testing sound discrimination.

Adding Text-Completion Items

The reading section of the Classic TOEIC Listening and Reading test included 40 incomplete sentence items measuring grammatical and lexical abilities in single-sentence contexts. To engage authentic interrelated language abilities more typical of the abilities necessary in communicative contexts in the real world, it was proposed that some longer contexts be created to contextualize items in passage format. Some of the items would require the examinee to connect information across different parts of a text in order to answer the question correctly. This design would allow examinees with greater communicative abilities to benefit from their understanding of the larger context.

Including Linked Passages

Another revision that was proposed to improve the measurement of reading abilities was the inclusion of some linked passages. These consisted of two related texts on a common topic. For example, the stimuli might consist of an e-mail exchange between two people, or an advertisement and a business letter on a common topic. Questions would be asked about both texts and some questions would require connecting information from both texts.

Evidence-Centered Design (ECD)

Concurrent with the analysis of the Classic TOEIC Listening and Reading test from the perspective of communicative competence, the redesign team began work on the second goal of the project: identifying and controlling major variables contributing to the difficulty of language tasks on the test, using ECD methodology.

ECD is a test-design process that begins with a discussion of the kinds of information it would be valuable to provide to test score users about the abilities examinees have or do not have, based on their test performance. Value is defined in terms of the information needed to support the decisions made by test users based on test scores. Test designers first decide what information about examinee abilities the test user needs; they then consider what evidence would support this information, and they determine how to obtain the evidence through tasks on the test. Test items are then designed so that evidence can be collected based on examinees' performance on the items. The evidence collected is what is then used to characterize each examinee's proficiency. Using this ECD approach, performance on test items is directly linked to evidence about language abilities. In addition to making explicit the rationale for selecting what abilities to measure, evaluating what evidence of these abilities to collect, and determining the types of items that will allow evidence to be collected, ECD also allows for an opportunity in the future to produce diagnostic proficiency information.

Stages of the Project

There were four stages to the project, which will be discussed here:

- Construct identification
- Prototyping
- Pilot testing
- Field testing

Construct Identification

In the construct identification stage, the Classic TOEIC Listening and Reading test forms were reviewed using an ECD approach. The design team considered the language abilities that examinees need in the real world, the abilities the Classic TOEIC Listening and Reading test measured, and ways to improve the precision of measurement so that more information about examinee proficiency could be provided. The goal was to connect performance on individual test items directly to the language abilities required for effective listening and reading. During this stage, test constructs were defined and articulated in terms of the abilities the test population requires, and proposed test tasks were linked to the required abilities.

At this initial stage of the project, the redesign team identified four useful types of information about examinees' listening abilities and ten types of information about their reading abilities, that the redesigned test could provide. In Evidence Centered Design, the information provided about examinees' abilities is written in the form of "claims" made about what test takers can or cannot do. The language abilities that underlie the claims identified for the Redesigned TOEIC Listening and Reading test were identified on the basis of current language theory and research. The redesign team looked at variables thought to affect the difficulty of performance on test items measuring these abilities. Table 2 lists the initial construct identification components.

TABLE 2

Initial Construct Identification Components

Claims about examinees' proficiency	Underlying abilities
Listening abilities	
Claim 1: Examinee can infer gist, purpose, and basic context based on information that is explicitly stated in spoken texts	The ability to understand and process multiple pieces of information The use of phonological, lexical, and grammatical abilities to understand the explicitly stated information needed to infer non-explicit information The ability to differentiate phoneme and stress (sound discrimination) patterns An understanding of morphological and syntactic markers
Claim 2: Examinee can understand details in talks and conversations on workplace and social topics and in descriptive sentences about photos	All Claim 1 abilities, plus The ability to remember important details from short and extended spoken texts
Claim 3: Examinee can understand interconnected ideas or information presented in monologic speech and conversations	All Claim 1 abilities, plus The ability to understand the interdependency of ideas and the connections among ideas in extended spoken texts
Claim 4: Examinee can make inferences based on information that is not explicitly stated in the spoken text	The ability to use phonological, lexical, grammatical, and semantic abilities to infer non-explicit information and feelings or emotions with regard to information in spoken texts The ability to understand rhetorical discourse features

Claims about examinees' proficiency	Underlying abilities
Reading abilities	
Claim 1: Examinee can understand specific (factual) information in tables and passages [understanding key/text information]	The ability to locate and match information in texts Knowledge of vocabulary An understanding of synonymous language (sentence-level paraphrase, vocabulary words and phrases) An understanding of syntactic and organizational structures
Claim 2: Examinee can connect information across multiple sentences in a single passage text	All of Claim 1 abilities, plus The ability to recognize lexical and grammatical relationships between and among ideas in a text
Claim 3: Examinee can connect information across two texts (e.g., an e-mail exchange, an exchange of letters)	All of Claim 1 abilities, plus The ability to recognize lexical and grammatical relationships between and among ideas across texts The ability to understand explicitly stated information to infer non-explicit information
Claim 4: Examinee can make inferences based on information that is explicitly stated in texts	All of Claim 1 abilities, plus The ability to recognize lexical and grammatical relationships between and among ideas in a text The ability to use explicitly stated information to make inferences
Claim 5: Examinee can make inferences based on information that is not explicitly stated in text (tone, attitude)	All of Claim 1 abilities, plus The ability to infer author's position with respect to information presented based on the manner of expression used and the denotative and connotative meaning of words chosen
Claim 6: Examinee can understand negative factual information	All of Claim 1 abilities + The ability to verify what information is true and what information is NOT true or NOT included in the passage, based on information that is explicitly stated in the passage

Claims about examinees' proficiency	Underlying abilities
Reading abilities	
Claim 7: Examinee can understand vocabulary	The ability to comprehend the meaning of individual words and phrases as used in the context of the passage
Claim 8: Examinee can understand grammar	Knowledge of the rules of grammar
Claim 9: Examinee can infer gist by connecting information across two texts	Sufficient processing ability to understand and remember multiple pieces of information The use of phonological, lexical, and grammatical abilities to understand explicitly stated information needed to infer non-explicit information
Claim 10: Examinee can infer gist by connecting pieces of information within a text	The ability to process and understand multiple pieces of information The use of phonological, lexical and grammatical abilities to understand explicitly stated information needed to infer non-explicit information

Note. Claims 5 and 9 were not represented in the field study.

The next step was to design prototypes of the proposed new and revised item types and to analyze and manipulate the characteristics of these items related to the underlying abilities needed to perform the tasks at different levels of difficulty.

Prototyping

In the second stage, prototyping, new item types were developed and tried out in testlets with two small groups — one on the Educational Testing Service site in Princeton, New Jersey, in the US, and the other in Japan. The goal of these tryouts was to collect information about the reactions of test users to the new item types and to ensure that the directions were clear enough to be used in subsequent pilot testing with large numbers of participants (see the appendix for copies of the debriefing protocols).

Generally the reaction of this small sample to the new item types was positive. Participants thought the testlets were a good measure of their abilities; they understood the directions, knew what they were supposed to do and did not find the new types of items to be culture bound. Some decisions were made on the basis of the small-scale prototype stage.

Varied accents did not pose problems for most candidates. A few lower-level people found one or another of the accents difficult to understand, but accent was just one variable in the difficulty of items for these participants.

More than half of all participants expressed interest in being able to take notes in the listening section and, although some people felt that the longer stimuli were too long, many people thought the length of the stimuli was appropriate. Longer stimuli, note taking, and varied accents were carried forward into the pilot testing stage to collect further information about the viability of the new item types.

Two alternative formats for the sentence completion passage-based items were presented in this small study. In one version, the answer choices were displayed under each space where an insertion was required. In the other version, the insertion spaces were numbered in the passage and the answer choices for all items were presented by number at the end of the passage. Most participants in the study expressed a preference for the former, the format that kept options closest to the insertion point in the passage — so this is the format that was used in the pilot testing stage.

Pilot Testing

With the third stage of the project, pilot testing, the emphasis of the redesign shifted from the theoretical to the empirical. The pilot was administered in February 2004, in Japan and Korea. Because a shorter version of the test was under consideration at the time, the full test pilot forms consisted of 140 items instead of 200 (70 items instead of 100 items per section). The sample size of 1,909 candidates was large enough to collect statistically significant information about the difficulty, reliability and scaling of the new item types. Examinees were administered a pilot form of the proposed redesign and an existing or the Classic TOEIC test form.

Two parallel pilot forms were created using the same stimulus materials, with some small changes in the listening stimuli. With the use of difficulty variables related to language abilities and test tasks that had been identified in the construct identification stage, each form was constructed with some items designed to be easy and others designed to be difficult. In one form, an item was designed to be relatively easy; in the other form, the same item was designed to be more difficult. The pilot forms were constructed this way because one of the goals of the pilot was to study difficulty variables. Another difference in the pilot test forms was that the order of the listening and reading sections was the reverse of that used in the Classic TOEIC test. Reading comprehension preceded listening comprehension because it was thought that this would help examinees access as much of their English vocabulary as possible before they were required to listen in English.

Table 3 compares the makeup of the reading sections of the Classic TOEIC Listening and Reading test and the Redesigned TOEIC test used in the pilot tests.

TABLE 3

Comparison of the Reading Sections of the Classic TOEIC Test and the Redesigned TOEIC Test

Reading section of the original TOEIC test (100 items)	Reading section of pilot the redesigned TOEIC test (70-items)
Incomplete sentences: 40 discrete items Error recognition: 20 discrete items Reading comprehension: 40 items; 12-15 stimuli with 2-5 questions each	Incomplete sentences: 20 discrete items Passage-based text-completion items: 10 items; 2 stimuli with 5 questions each Reading comprehension: 29 items; 5-6 stimuli with 5-6 questions each Tables: 6 items; 2 tables with 3 questions each Linked passage sets: 5 items; 1 stimulus with 5 questions

New Reading Item Types

All the item types that had been prototyped with small groups of participants in the second stage were administered in the pilot test forms as well. Two types of table-formatted stimuli were included; one type contained more language and the other contained more symbols. Two linked passage sets were also included, as were two text-completion passages.

The pilot forms were designed to measure the reading claims identified in the construct identification stage:

- Examinees can understand specific (factual) information in tables and passages.
- Examinees can connect information across sentences.
- Examinees can connect information across passages.
- Examinees can make inferences.
- Examinees can understand vocabulary.
- Examinees can understand grammar.

The importance of these high-level abilities is supported in the theoretical and research literature.

To address the project goal of providing more specific information about examinees' strengths and weaknesses in English, each item in both forms was linked to the ability (of the six abilities listed above) that an examinee theoretically would need in order to respond correctly to the question. In addition to linking each item to a specific ability, a number of item-specific difficulty variables were coded for each item. These were the item characteristics believed to make the item easier or more difficult than its "twin" in the other pilot form.

Table 4 compares the makeup of the listening sections of the Classic TOEIC Listening and Reading test and the Redesigned TOEIC Listening and Reading test used in the pilots.

TABLE 4

Comparison of the Listening Sections of the Classic TOEIC Test and the Redesigned TOEIC Test

Listening section of the Classic TOEIC Listening and Reading test (100 items)	Listening section of the Redesigned TOEIC test Listening and Reading in pilot tests (70-items)
Photographs: 20 discrete items	Photographs: 10 discrete items
Question-response: 30 discrete items (3 choices—all other item types in test have 4 choices)	Question-Response: 20 discrete items (3 choices—all other item types in test have 4 choices)
Short conversations: 30 discrete items	Conversations: 20 items 5-6 stimuli with 3-4 questions each
Short talks: 20 items; 7-9 stimuli with 2-4 items each	Talks: 20 items 5-6 stimuli with 3-4 questions each

New Listening Item Types

All the item types that had been prototyped were administered in the pilots as well. Major differences in the listening pilot included longer oral stimuli in the conversation and talk parts of the test, the use of item sets (rather than discrete items) in the conversations, questions that are spoken as well as written in the test book; and the inclusion of a greater variety of accents in the recorded listening material.

The pilot listening forms were designed to measure the listening claims identified in the construct identification stage:

- Examinees can differentiate phoneme and stress patterns (sound discrimination).
- Examinees can infer gist, purpose, and basic context.
- Examinees can understand details.
- Examinees can connect information and ideas.
- Examinees can make inferences based on information that is not explicitly stated.

As was the case with items in the reading section of the pilot test, each of the items in the listening section was linked to one of these abilities and coded for difficulty.

From a content and test design point of view, the new item types in the pilot were successful. Hypotheses about difficulty variables were largely confirmed. Items that were thought to be more difficult because of their difficulty variables generally were more difficult than items thought to be easier because of their difficulty variables. Item analysis supported hypotheses about many of the variables predicted to influence difficulty. It was shown to be possible to predict the *direction* of difficulty in relative terms — easier or more difficult — related to the task variables used in designing the items. Whether the difficulty of test items could be predicted with any degree of precision was a question for the final stage of the project, as was the question of how comparable in difficulty the proposed redesigned test would be to the Classic TOEIC Listening and Reading test and whether results could be reported using the existing TOEIC scale.

The reliability of the pilot test forms was lower than that of the Classic TOEIC Listening and Reading test. There were two reasons for this. Items had been designed to try out variables affecting difficulty, so a greater percentage of items were at the extremes—more very easy items and more very difficult items—and items at the extremes do not discriminate as well. Another reason for the lower reliability was that the number of items in each of the pilot forms had been reduced from 200 to 140. For the TOEIC field study, the first issue would be addressed by developing items targeted at difficulty levels typical of the Classic TOEIC Listening and Reading test. The second issue would be addressed by creating a full 200-item test.

Field Study

The field study was the fourth and final stage of the project, and it was administered in November 2004 (see Compendium Study 3.1). It was a large-scale administration of two parallel forms that differed in content but were designed to the same test specifications. Again, an original form of the Classic TOEIC Listening and Reading test was also administered. One purpose of the field study was to evaluate the psychometric quality of the proposed TOEIC test redesign for the TOEIC population. Another was to evaluate the usefulness of item coding related to the claims, abilities and difficulty variables.

Between the pilot-testing stage and the field study, a decision had been reached to retain the 200-item structure of the Classic TOEIC Listening and Reading test, rather than to shorten the test as had been proposed at the time of the pilot testing. Therefore, the field study forms were 200 items in length. Table 5 summarizes the final design.

TABLE 5

Final Design of the Redesigned TOEIC Test in the Pilot Tests

Section 1 Listening Comprehension (Items 1-100)	
Items 1-10	Photographs (10 items)
Items 11-40	Question-responses (30 items)
Items 41-100	Conversations and talks (60 items)
Conversations with speaker exchange pattern of ABA or ABAB; 10 stimuli with 3 questions about each (30 items) Talks 10 stimuli with 3 questions about each (30 items)	

Section 2 Reading Comprehension (Items 101-200)	
Items 101-140	Incomplete sentences (40 items; 20 vocabulary & 20 grammar)
	Items 141-152 Text-completion sets (12 items total) 3 sets with 4 questions each (subsequently revised to 4 sets with 3 questions each)
	Items 153-180 Reading sets, including tables or charts (28 items total) 9 sets with 2-4 items each
	Items 181-200 Double passage sets (20 items total) 4 sets with 5 questions each

In the pilot tests, the order of the sections was the reverse of that in the Classic TOEIC Listening and Reading test. For the field study, the original order (listening section followed by reading section) was restored for practical reasons related to the ease of administering the test.

Items in both forms of the field study were coded for the listening and reading claims and abilities identified in the construct identification stage. However, reading claims 5 and 9 were omitted because it had not been possible in the pilot testing to generate sufficient numbers of these items for accurate predictions of examinee abilities.

Each item was also coded for the difficulty variables that had appeared to be most relevant in the pilot-testing stage, and these variables were used to design sets of items predicted to be of parallel difficulty in the two forms.

The field study was intended to answer the following questions:

- Could the existing TOEIC scale be used to report performance on the proposed redesign of the Classic TOEIC Listening and Reading test?
- Could the new TOEIC items be designed to replicate the overall difficulty and reliability of the Classic TOEIC Listening and Reading test?
- Could the listening and reading claims and underlying abilities (those identified in the construct identification stage and shown to be most relevant in the pilot testing) be used to provide improved feedback to the test score users about the language proficiency of the TOEIC examinees?

The first and second questions have been answered in the affirmative (see Compendium Study 4.1). The Redesigned TOEIC items were found to be comparable to the Classic TOEIC items in difficulty and reliability, and the Redesigned TOEIC test fits the existing TOEIC scale. The third question — whether the claims and underlying abilities associated with the Redesigned TOEIC items can be used to provide improved feedback to test score users — has also been answered in the affirmative. An analysis of the listening and reading claims was carried out to determine the extent to which the claims and abilities could be used to provide additional score report information. In some cases, too few items had been included to support reliable feedback for that claim. Claim 2 of the listening claims (Understanding Details) had more than enough items associated with it, but there were too few items associated with claims 3 and 4. This imbalance was addressed by splitting claims 1 and 2 (Inferring Gist and Understanding Details) into four claims by adding the component of length to further differentiate the abilities needed to infer gist and understand details. Table 6 compares the listening claims that resulted to the listening claims originally identified in the ECD stage. The original claims 3 and 4 are now accounted for as part of the new claims 3 and 4, resulting in the same number of claims but a better distribution of items in each claim category.

TABLE 6

Comparison of Claims About the Listening Section

Classic ECD listening claims	Listening claims post field study
Claim 1: Examinee can infer gist, purpose and basic context based on information that is explicitly stated in spoken texts Claim 2: Examinee can understand details in conversations on workplace and social topics and in descriptive sentences about photos Claim 3: Examinee can understand complex interconnected ideas or information presented in monologic speech and conversations Claim 4: Examinee can make inferences based on information that is not explicitly stated in the spoken text	Claim 1: Examinee can infer gist, purpose and basic context based on information that is explicitly stated in short spoken texts Claim 2: Examinee can infer gist, purpose and basic context based on information that is explicitly stated in extended spoken texts Claim 3: Examinee can understand details in short spoken texts Claim 4: Examinee can understand details in extended spoken texts

In the reading section, claims for items with too few exemplars in the field study were merged with claims for items requiring very similar or identical underlying abilities. Table 7 compares the resulting claims about the reading section to the claims about the reading section originally identified in the ECD stage.

TABLE 7

Comparison of Claims About the Reading Section

Classic ECD reading claims	Reading claims post field study
Claim 1: Understanding specific (factual) information in tables and passages [understanding key/text information]	Claim 1 & Claim 6 combined: Understanding specific (factual) information in tables and passages, including negative factual information (understanding key/text information)
Claim 2: Connecting information across multiple sentences in a single passage text	Claim 2 & Claim 3 combined: Connecting information across multiple sentences (across a single text, across 2 texts)
Claim 3: Connecting information across two texts (e.g., an E-mail exchange, an exchange of letters)	
Claim 4: Making inferences based on information that is explicitly stated in texts	Claim 4 & 10 combined: Making inferences based on information that is explicitly stated in texts (within text local inferences and across text gist)
Claim 5: Making inferences based on information that is not explicitly stated in text (tone, attitude)	
Claim 6: Understanding negative factual information	
Claim 7: Understanding vocabulary	Claim 7: Understanding vocabulary
Claim 8: Understanding grammar	Claim 8: Understanding grammar
Claim 9: Inferring gist by connecting information across two texts	
Claim 10: Inferring gist by connecting information across a text	

Note. Claims 5 and 9 were not represented in the field study.

After these revisions were implemented, a second analysis of the claims was run on all items in both forms of the field study, and a high degree of reliability for this type of analysis was achieved for all claims.

Next Steps

With decisions about test design finalized and test specifications in place, next steps were to determine exactly what types of information would be provided to test score users and how this type of information could be most effectively presented. Tying items on each test form to specific claims about examinee abilities would allow examinees to receive a percent-correct score for each of several language abilities. In addition, a scale anchoring study was instituted so that scaled score points on the test scale could be aligned with descriptions of language skill areas (see Chapter 5).

References

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). TOEFL 2000 listening framework: A working paper (TOEFL Monograph No. MS-19). Princeton, NJ: ETS.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). TOEFL 2000 reading framework: A working paper (TOEFL Monograph Series No. MS-17). Princeton, NJ: ETS.

Appendix

Prototyping Stage Debriefing Protocol

Questions interviewers asked about each item in the listening and reading sections

- How did you decide which answer was correct?
- If you weren't sure which answer was correct, how did you select (whatever they picked)?
- What method did you use to reach your decision?
- Why did you choose this answer over the other choices? [In reading, we would like to know what they thought words, phrases, or sentences meant. In listening, we would like to know what they thought they heard.]
- Was this item easy or difficult for you to answer, and why?

Questions interviewers asked about each set in the reading section

- Are you familiar with this question type from previous learning/tests? YES NO
- Were the directions for these questions clear? YES NO
- Did you understand what you were supposed to do? YES NO
- Was any vocabulary (wording) unclear or confusing? YES NO
- Was any grammar unclear or confusing? YES NO
- Culture-bound: Is this kind of question (passage, conversation, etc.) familiar to you?
  YES NO

Questions interviewers asked at the end of the reading section

- Was this test a fair measure of your abilities? YES NO
- Did this test give you a fair opportunity to show your skill in reading English? YES NO
- Did you like some types of questions more or less than other types? If so, which ones? Why?
- What were the most difficult items in the reading test for you (or select an item or two that was particularly difficult)? Why were these difficult?
  (possible prompts: vocabulary, length, cloze format or regular m.c. format)
- What were the easiest items in the reading test for you (or select an item or two that was particularly easy)? Why were these easy?
  (possible prompts: vocabulary, length, cloze format or regular m.c. format)
- Did you think any of the question types were not a good measure of your reading proficiency?
  [Check here about the graphic table and the cloze formats]
- Which of the two text-completion item formats do you prefer? [We are presenting two styles] Why?
- What did you think about the graph/table items? Did you prefer one type over the other? Why?

Questions interviewers asked about each set in the listening section*

- Are you familiar with this question type from previous learning/tests? YES NO
- Were the directions for these questions clear? YES NO
- Did you understand what you were supposed to do? YES NO
- Was any vocabulary (wording) unclear or confusing? YES NO
- Was any grammar unclear or confusing? YES NO
- Was the length of the stimulus (passage, short talk, long talk, etc.) too long? YES NO
- Memory load: Were you able to remember everything presented to you (passage, short talk, long talk, etc.) in order to answer the first question, or was there too much information to remember?
- Were you able to remember everything presented to you (passage, short talk, long talk, etc.) in order to answer the second question, or was there too much information to remember?
- Culture-bound: Is this kind of question (passage, conversation, etc.) familiar to you?
  YES NO

For photographs, only the first 5 questions were asked.

Questions asked at the end of the listening section

- Was this test a fair measure of your abilities? YES NO
- Did this test give you a fair opportunity to show your skill in listening to English? YES NO
- Would you like to be able to take notes while you are taking a test like this? YES NO
- Did you think any of the speakers were easier or harder to understand? If so, which one(s) (prompts: why? accent? speed of delivery?) [There were four accents: US woman/ British woman/Canadian man/Australian man.]
- Did you like some types of questions more or less than others? If so, which ones? Why?
- What were the most difficult items in the listening test for you (or select an item or two that was particularly difficult) Why? (possible prompts: vocabulary, accent, speed of delivery, length)
- What were the easiest items in the listening test for you (or select an item or two that was particularly easy) Why? (possible prompts: vocabulary, accent, speed of delivery, length)
- Did you have enough time to answer or did you feel hurried? [This might apply more to listening, since reading is not paced.]