ETS®

TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

Evidence-Centered Design: The TOEIC® Speaking and Writing Tests

Susan Hines

January 2010

Based on preliminary market data collected by ETS in 2004 from the TOEIC® test score users (e.g., companies, educational institutions, and universities) and potential test takers in Japan, Korea and Mexico, it was clear that score users were interested in an assessment that would give them information about test takers' speaking and writing abilities. The team of test developers, market analysts and statisticians responsible for this exploratory investigation created initial test prototypes based on other modality-relevant tests created by ETS including the *Test of Professional English* (TOPE™), the Test of Spoken English™ (TSE®), and the TOEFL iBT™ test. This small scale pilot, in addition to rater evaluation and scoring guide analysis, provided the foundation for a full-scale test design process and field study. In August 2005, ETS staff met with ETS Preferred Network members from Korea (YBM/Sisa) and Japan (IIBC). Part of this meeting was devoted to outlining and confirming the business requirements for the addition of speaking and writing tests to the TOEIC family of products. The test designers emerged from this meeting with the following business requirements which had immediate relevance to test design:

- The tests should be linear and delivered by computer with perhaps a paper option for writing.
- Each test should discriminate across a wide range of abilities; particularly they should discriminate among candidates of relatively low ability (as low as test takers with TOEIC Listening and Reading test combined scores of 400, who are in the bottom quintal of TOEIC test takers).
- Each test should separate candidates into as many levels as possible.
- Combined testing time for the TOEIC Speaking and Writing tests should be approximately 90 minutes.

In addition to considering the requirements above, the test designers recognized that test security would also be a serious issue. Because of test security concerns, the test design team at ETS knew that multiple parallel test forms, with minimal reuse of items, would be necessary. Based on ETS' experience with developing test forms for the TOEIC Listening and Reading test and the TOEFL iBT test, it was anticipated that a range of 50–70 unique test forms would need to be created every year to be administered across potentially hundreds of testing sessions throughout the year. To make the development of many parallel forms possible, the TOEIC Speaking and Writing tests had to include detailed task specifications that could be clearly communicated to a large number of test writers.

Another major issue for the test designers to consider was how to maximize rater accuracy and efficiency—two factors which can often be at odds with one another. Fair and accurate scoring was of supreme concern for the test designers at ETS. However, they also understood that score users needed fast score turnaround for timely decision-making. In order to facilitate timely decision-making processes for score users and test takers, the test designers were aware that they needed to consider the difficulty of scoring each task in the tests.

After considering the business requirements and major design issues, a final design process was needed to produce task specifications and a test blueprint that would ultimately support the generalization of test scores from any individual set of test tasks to performance on actual tasks required in the workplace. This paper describes task design analyses as practical processes for tackling these issues and task shells as documents that show the design decisions made to construct the final test blueprints.

Evidence-Centered Design as the Basis for Task Design Analysis

The processes and concepts drawn from discussions of evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2002, 2003) served as the foundation for task design analysis. Examples of how task design analysis was used to create task specifications for speaking and writing tasks are provided in this paper. In ECD, task specifications are entered into a framework called a task shell, which is described and illustrated with an example of a speaking task.

Evidence-Centered Design

ECD can be viewed as a methodology that comprises best practices for the creation and on-going development of an assessment tool. It clarifies what is being measured by a test and supports inferences made on the basis of evidence derived from the test. ECD was useful for the design of the TOEIC® Speaking and Writing tests because it oriented the test designers toward competencies and tasks, and the relationships between them, during the test design process. ECD as described by Mislevy et al. (2002, 2003) and summarized in Table 1, systematizes test design by specifying a four-stage process consisting of: domain analysis, domain modeling, construction of a conceptual assessment framework, and deployment of an operational assessment. These stages concretize what we want to be able to say about test takers based on observations we make on their performance on the test tasks.

TABLE 1

Mislevy's Four Stage Evidence-Centered Design Process for Test D

Stage	Process	Component	Definition of component
1. Domain analysis	Preliminary synthesis of what is known about what is to be assessed	No specific components specified although useful categories enumerated	NA
2. Domain modeling	Incorporation of information from stage one into three components; sketch of potential variables and substantive relationships	Proficiency paradigm	Substantive construct expressed as claims
		Evidence paradigm	Observations required to support claims
		Task paradigm	Types of situations that provide opportunities for test takers to show evidence of their proficiencies
3. Construction of conceptual assessment framework	Development of a final blueprint; provide technical detail required for implementation including statistical models, rubrics, specifications and operational requirements	Student model	Statistical characterization of the abilities to be assessed
		Evidence model	1. Rules for scoring test tasks
			2. Rules for updating variables in the student model
		Task model	Detailed description of assessment tasks
		Presentation model	Specification of how the assessment elements will look during testing
		Assembly model	Specification of the mix of tasks on a test for a particular student
4. Deployment of operational assessment	Construction of the operational delivery system	Presentation	Presentation, interaction and response capture
		Response scoring	Evaluation of response; task level scoring
		Summary scoring	Computation of test score; test level scoring
		Activity selection	Determine what to do next

The first stage of test design, domain analysis, consists of a preliminary synthesis of what is known about the field to be assessed and focuses on questions such as the following:

- What are the skills, knowledge and abilities important for success in the field?
- What are the real world situations in which we can see people using the kinds of knowledge we care about?
- What are the important features of these situations?
- What theoretical perspectives have been proposed to explain performance?

While this information was not originally created or organized for the purposes of generating an assessment tool, it provides a foundation on which an assessment tool can be further developed.

In the second stage, domain modeling, the information collected in the first stage is refined as it is incorporated into three interrelated components or structures that will guide the development of the assessment:

1. Proficiency Paradigm—What substantive claims will be made about test takers' abilities or competencies?

2. Evidence Paradigm—What observable features in test takers' performances would provide data to support these claims?

3. Task Paradigm—What kinds of tasks provide an opportunity for test takers to demonstrate evidence of their proficiencies?

Stage three, the Conceptual Assessment Framework (CAF), adds technical detail to the sketch obtained in the domain analysis stage. The CAF is made up of five models. The student model is a statistical characterization of test takers such as their locations on a continuous, or unidimensional scale, or at a specific level on an ordered categorical scale. The evidence model has two subcomponents: The evaluation component prescribes the rules for scoring test tasks and the measurement component contains the statistical machinery to be used to accumulate data across tasks to update the student model. The task model provides a detailed description of the characteristics of test tasks. The task model, in many ways akin to task specifications, provides the direction needed to generate multiple exemplars of a particular task type. For linear tests, the assembly model, which stipulates the mix of tasks to be presented to test takers, corresponds to the test blueprint. The test blueprint provides a template that details requirements that must be met for each test form, such as the number tasks of each type to be presented, the content to be covered, and the order of the tasks. Finally, the presentation model lays out the formatting specifications for a test.

The final stage of ECD, an operational field study, consists of a four-process delivery system. This includes (a) presentation of information, interaction between the information and the test taker, and the capture of the test takers' response (b) scoring of the response (c) summarization of the scores across several responses and (d) decisions about possible future steps.

ECD provides a framework to formalize and document traditional test design processes in greater detail and to more clearly articulate the connections between elements of the test design. Because ETS already had experience using ECD to create the final blueprint for the TOEFL iBT test, as well as experience gained from using ECD principles to shape the redesign of the TOEIC Listening and Reading test, the TOEIC Speaking and Writing tests design team was able to build on this pre-existing body of work.

Task Design Analysis

Task design analysis (TDA) was conducted by following six steps. These steps were formulated based upon the general principles of ECD. In particular, the six steps of TDA drew upon the first two stages of ECD, domain analysis and domain modeling. The correspondences between the components of domain analysis and domain modeling that were relevant to each step of TDA are outlined in Table 2.

TABLE 2

Steps Carried Out in Task Design Analysis Guided by Aspects of Evidence-Centered Design

Step in task design analysis	Component from evidence-centered design	Stage of evidence-centered design
Reviewing prior theory and research pertaining to testing issues	No specific components defined	Stage 1—Domain analysis: preliminary synthesis of what is known about what is to be assessed
Articulating claims about test takers' language proficiency in all modalities and stating more detailed claims as subclaims	Proficiency paradigm	Stage 2—Domain modeling: incorporation of information from stage 1 into three components; sketch of potential variables and substantive relationships
Listing sources of evidence for each claim	Evidence paradigm	
Listing real world tasks for which test takers can provide relevant evidence	Task paradigm	
Identifying characteristics that could affect task difficulty	Task paradigm	
Identifying criteria for evaluating performance on the tasks	Task paradigm	

The first step in TDA is reviewing prior theory and research pertaining to the testing issues. The design team, consisting of assessment specialists at ETS, for the TOEIC Speaking and Writing tests accomplished the first step by reviewing the TOEFL test framework papers (Butler, Eignor, Jones, McNamara, & Suomi, 2000; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000). They read and discussed these papers in addition to the other materials about design and research that had been produced regarding workplace and general English tests (e.g., redesign of the TOEIC Listening and Reading test, TOPE, and TSE). This review of prior work provided them with ideas about potential test tasks that would be useful for the subsequent five steps.

The second step of TDA, articulating claims, is a means of specifying the proficiency paradigm for the test. The proficiency paradigm is specified as the substantive ability construct that the test is intended to measure. Following Mislevy et al. (2002, 2003), such constructs are expressed as claims that one would like to make about test takers. Using the speaking measure as an example, the construct of the ability to communicate successfully in everyday life and the international workplace context would be expressed as a claim such as, "The test taker can communicate effectively by communicating in spoken English to function effectively in the context of a global workplace and everyday life". Such a general claim can be specified further through the development of subclaims. A subclaim for the speaking assessment, therefore, provides a means of articulating a more specific construct such as the ability to speak about something in a particular context: The test taker can select appropriate language to carry out routine social and occupational interactions (such as giving and receiving directions; asking for and giving information; asking for and providing clarification; making purchases; greetings and introductions; etc.).

The third step, listing sources of evidence for claims, refers to the process of defining an evidence paradigm. The evidence paradigm characterizes the observations that are required to support claims by listing as many sources of evidence as possible for each of the claims. Continuing with the example of the speaking measure, the following aspects of the test takers' responses were identified as the relevant evidence in the spoken response: task appropriateness, delivery, relevant vocabulary and use of structures.

Steps 4–6 of TDA define a task paradigm by listing real world tasks in which test takers can provide relevant evidence, identifying task characteristics that might affect difficulty, and establishing criteria for evaluating performance. For example, real world tasks involving speaking skills consisted of those requiring test takers to ask and respond to questions based on written information in a workplace setting, participate in a discussion that requires problem-solving, and exchange information one-on-one with colleagues, customers or acquaintances. Task characteristics potentially affecting difficulty included the characteristics of the reading and listening material and the nature of their connections with each other. Important features for evaluating speaking performance on these types of tasks include: the range and complexity of vocabulary and structures; clarity and pace of speech; coherence and cohesion; progression of ideas within response; relevance and thoroughness of the content of the response. Table 3 summarizes the outcomes from the six-step TDA conducted for the speaking measure.

TABLE 3

Example Task Design Analysis for Speaking

Step in task design analysis	Outcome for the speaking test
Reviewing previous research and other relevant assessments	Ideas about language proficiency and potential test tasks
Articulating claims and subclaims	Claim: The test taker is able to communicate in spoken English which is needed to function effectively in the context of a global workplace. Subclaims: The test taker can generate language intelligible to native and proficient nonnative English speakers. The test-taker can select appropriate language to carry out routine social and occupational interactions (such as giving and receiving directions; asking for and giving information; asking for and providing clarification; making purchases; greetings and introductions; etc.). The test-taker can create connected, sustained discourse appropriate to the typical workplace.
Listing sources of evidence	Task appropriateness, delivery, relevant vocabulary and use of structures
Listing real world tasks in which test takers can provide relevant evidence	Asking and responding to questions based on written information in a workplace setting, participating in a discussion that requires problem solving, and exchanging information one-on-one with colleagues, customers, or acquaintances
Identifying aspects of situations that would affect their difficulty	Characteristics of reading and listening material; the nature of their connections to each other (referring to Subclaim 2)
Identifying criteria for evaluating performance on the tasks	Range and complexity of vocabulary and structures; clarity and pace of speech; coherence and cohesion; progression of ideas in response; relevance and thoroughness of the content of the response

The outcome of TDA brings the test design team closer to defining the test tasks by compiling the necessary information. This information, which is exemplified for the speaking test in Table 3, is used to construct the task shells. These task shells are then used in turn to generate possible test tasks.

Task Shells

A task shell is a template for generating parallel items or test tasks. Task shells are composed of two primary parts: a summary of what the task is intended to measure and a task model.

The summary of what is being measured is shown in Table 4, which contains two entries in the first column. These entries provide the over-arching ideas that drive development of more detailed test specifications. The first entry states the claim that the test design team intends to make about the test taker on the basis of performance on the task. This claim comes directly from Step 2 of the TDA. The second entry includes the observable behaviors of the test taker that are to be used to provide evidence for support of the claim. This comes from Step 3 of the TDA.

TABLE 4

Components of a Task Shell

What is being measured?	Task model			
	Fixed elements	Variable elements	Rubric	Variants
Claim: **Statement that one would like to make about the test taker on the basis of test results** *Measurement:* **Aspects of the test taker's response that are the focus of evaluation**	Aspects of this class of task that remain constant across tasks • Nature of the task: Description of what students are asked to do and materials that they are given • Order of item elements: Sequence and timing of task components	Aspects of this class of task that can be changed across tasks	Procedures and criteria for evaluation of learners' responses	Descriptions of example tasks

The task model in the task shell comes from Steps 4–6 in the TDA, as shown in Table 4. The task model consists of four parts: fixed elements, variable elements, the rubric or scoring guides, and variants. The fixed elements refer to the aspects of a class of tasks that remain constant across different instances of that task. The features specified are what the test takers are asked to do to complete the task, and the sequence and timing of the task components. Examples of fixed features for the TOEIC Speaking test include exact length of stimulus materials; exact timing for stimulus, preparation, and response; and exact specification of the nature of the prompt. The variable elements refer to the aspects of the class of tasks that can be changed across different instances of that task, such as the types of texts or topics. The set of possible varieties of these elements is also included. The scoring guide specifies the procedures and criteria for evaluation of the test takers' responses. The variants define the range of tasks and specific topic areas or types of texts or examples of tasks defined by the task shell.

An example of a task shell for a speaking task is shown in Table 5. In the first column, a claim is specified based on those claims identified for speaking during Step 2 of the TDA (shown in Table 2). The claim in this example, “test takers can produce language that is intelligible to native and proficient non-native English speakers”, reflects Subclaim 1 in Table 3. The phrase above the claim in Table 5, “spoken ability producing the sounds and rhythms of English based on a text that would authentically be read aloud”, summarizes what the tasks developed from this task shell are intended to measure. Under the claim, the aspects of the test takers’ responses that are being measured are listed. The task model is completed with the specifications for the fixed elements, variable elements, rubric and list of variants. While the rubric is part of the task model, the amount of information contained within it is better represented in a separate document.

TABLE 5

An Example of a Task Shell for a Read a Text Aloud Speaking Task

What is being measured?	Task model			
	Fixed elements	Variable elements	Rubric	Variants
Spoken ability producing the sounds and rhythms of English based on a text that would authentically be read aloud Claim: Test takers can produce language that is intelligible to native and proficient non-native English speakers. Measurement: Analytic evaluation of: 4. Pronunciation of sounds 5. Intonation & Stress of sentences	1. Nature of the task Demonstrate ability to read two short texts aloud. Features of the texts to be read aloud: • 40–60 words • At least one complex sentence • A list of at least three elements • A transition • Context that is both? missing copy or combine with bullets below? • Accessible to beginning-level language learners • Text that would authentically be read aloud 2. Order of item elements • Task-specific directions will be both spoken and written: In this part of the test, you will read aloud the text on the screen. You will have 45 seconds to prepare. Then you will have 45 seconds to read the text aloud. • Preparation time: 45 seconds • Response time: 45 seconds	1. Type of texts 2. Topic of texts	See Table 7.12	Including, but not limited to: • Advertisement • Announcement • Broadcast • Directions/instructions • Excerpt from a talk • Introduction to an interview • Introduction • Recorded information • Recorded message • Report • Speech • Talk • Tour information • Traffic report • Weather report • Activities • Entertainment • Health • Housing • News • Shopping • Travel

The task shell shown in Table 1 was used to develop the task shown in Table 5. The task requires the test taker to read aloud two texts whose characteristics are specified under the fixed elements in the task model. In the 2006 field study, a single, longer text was piloted. However, in order for ETS to reliably report analytic scores, a decision was made to include two shorter texts. As the task model indicates, the topic and the types of texts can vary; therefore the example of recorded message as a text type is only one possible example of the type of text that could appear in such a task. The lists of variants are not all inclusive, but they provide further specification of the many possible sources of authentic texts and topics for this task type.

Once a shell and some sample tasks were created, the shell was evaluated by assessment specialists according to mutually agreed upon criteria based on content expertise and business requirements. A proposed task shell had to provide evidence for a subclaim, sustain the production of many variations of the same kind of task, contribute to content representation of the domain, and not be easily undermined by inappropriate test-taking strategies.

Task shells were developed for each of the subclaims shown in Table 7 that define the aspects of speaking to be measured on the speaking test. Similarly, TDA was conducted for the writing test and task shells were developed based on the claims and subclaims. These analyses resulted in the final test blueprint and the specifications for each of the measures that are summarized in the next section.

The Final Blueprint and Specifications for the Measures

The TDA process resulted in few changes to the speaking and writing measures after the field study. Again, this is due to the pre-existing body of work already documented in other relevant ETS tests. These changes are evident in the contrasts between the 2006 field study blueprint with the final blueprint.

Modifications to the Test Blueprint

Table 6 compares the 2006 field study blueprint with the final blueprint. There were very few changes made to the overall test format based on the results from the field study. Most of the hypotheses the test designers had put forward were confirmed in the results. Because of the previous experience in applying ECD to the TOEFL iBT test and the redesign of the TOEIC Listening and Reading tests, the design team for the TOEIC Speaking and Writing tests was able to make use of much of what had been learned during that process.

Of the few changes made, most were made to the speaking test. One of the value-added features built into the TOEIC Speaking test was the inclusion of descriptive analytic feedback on pronunciation and intonation and stress on a test taker's score report. This was included so that even the very lowest ability test takers would receive some information about their speaking ability on which to build for continued language study. In order to meet ETS standards for fair and reliable score reporting for analytic scores, it was decided that two shorter read-aloud texts would replace the original longer text, so there would be more data to support analytic score reporting. The only other change worth noting is related to the omission in operational testing of a task type that was tried out during the field study. The task type required a test taker to read a table where two products were compared side-by-side and make a recommendation using reasons and rationale. This task type exhibited different enough statistical results that it was decided that this task type was not as comparable to the other task types supporting the most difficult subclaim, Subclaim 3.

The writing test had almost no changes made to it. The test designers solidified test specifications at the form level after the field study, which included:

1. Balancing each form with a combination of easy and difficult picture sentence tasks represented in questions 1–5.

2. Specifying that each form would contain one response to a written request task where the test taker would respond as her/himself and one task where he/she would be required to role-play.

TABLE 6

Comparison of 2006 Blueprint and the Final Blueprint for the TOEIC Speaking and Writing tests

2006 field study blueprint		Final blueprint—TOEIC Speaking and Writing	
Stimulus	**Items per form**	**Stimulus**	**Items per form**
Section 1. integrated speaking – 10 items		Section 1. integrated speaking – 11 items	
Read a text aloud	1 text	Read a text aloud	2 texts
Describe a picture	1 photo	Describe a picture	1 photo
Respond to questions (listening)	1 set with 3 questions - 2 short questions - 1 longer question	Respond to questions (listening)	1 set with 3 questions - 2 short questions - 1 longer question
Respond to questions based on written information (reading/ listening)	1 set with 3 questions - 2 basic information questions - 1 question that requires summary	Respond to questions based on written text (reading/listening)	1 set with 3 questions - 2 basic information questions - 1 question that requires summary
Extended listening in form of a voice mail OR written comparison chart	1	Extended listening in form of a voice mail	1
Express an opinion	1 open or paired choice	Express an opinion	1 paired choice

2006 field study blueprint		Final blueprint—TOEIC Speaking and Writing	
Stimulus	Items per form	Stimulus	Items per form
Section 2. writing		Section 2. writing	
Picture sentences	5 total including combination of easy, medium and difficult	Picture sentences	5 total including mix of easy and hard word combinations
Respond to an e-mail	2	Respond to an e-mail	2 including 1 that allows test taker to respond as themselves and 1 where he/she must role play
Express an opinion	1 open or paired choice	Express an opinion	1 paired choice
Total test time ~ 1 hour 30 minutes		Total test time ~ 1 hour 30 minutes	

Summary of Specifications for the Measures

The outcome of ECD was a set of task specifications for each of the measures. These specifications included an overall claim and subclaims about what each measure is intended to assess. Linked to each of the subclaims are task model components that describe the nature of the task, response type, scoring guides, number of questions, the nature of the stimulus information, and task or section timing. The summary that follows for each measure describes the total number of questions and stimulus materials included on an operational form.

Speaking- The results of the TDA for the speaking measure are summarized in Table 7. As described earlier, three subclaims were identified that provided support for the overall claim that a test taker can communicate effectively by using English to communicate meaningfully in the workplace and everyday life. These subclaims were ordered hierarchically with the assumption that those who perform well on the most difficult tasks will also perform well on the intermediate- and beginning-level tasks. Task types that would provide evidence related to each of the three subclaims were defined, including two read aloud texts and one describe a picture task supporting Subclaim 1, two sets of questions requiring a combination of listening/reading or reading/listening/speaking skills supporting Subclaim 2, and two difficult tasks requiring sustained, coherent discourse supporting Subclaim 3. The stimulus materials for the speaking tasks represented appropriate contexts and language for the claim for which they were providing evidence.

TABLE 7

Summary of Specifications for Speaking Measure of the TOEIC Test

Speaking claim	Test taker can communicate in spoken English to function effectively in the context of a global workplace.					
Subclaims	Test taker can generate language intelligible to native and proficient nonnative English speakers.		Test taker can select appropriate language to carry out routine social and occupational interactions (such as giving and receiving directions; asking for and giving information; asking for and providing clarification; making purchases; greetings and introductions; etc.).		Test taker can create connected, sustained discourse appropriate to the typical workplace.	
Nature of speaking task	Read a text aloud	Describe a picture	Respond to short questions based on personal experience in the context of a telephone market survey	Respond to short questions based on information from a written schedule/ agenda	Propose a solution based on a problematic situation stated in the context of a voice mail message	Describe and support opinion with respect to a given pair of behaviors or courses of action
Scoring rubric	Analytic 0–3	Independent 0–3	Integrated 0–3	Integrated 0–3	Integrated 0–5	Independent 0–5
Number of questions	2	1	3	3	1	1

Speaking claim	**Test taker can communicate in spoken English to function effectively in the context of a global workplace.**					
Nature of stimulus material	Reading text that contains: complex sentence list of three items transition 40–60 words Text must be accessible to low-level speakers	Photograph that represents high frequency vocabulary or activities	Listening stimuli made up of three, short, related questions that are both seen and heard by the candidate; lead-in sets context for the topic of the questions; voices represent English speaking voices from the US, Australia, Britain and Canada	Reading passage: telegraphic text in the form of an agenda or schedule (65–75 words; 12 lines max.) Listening stimulus: Three short questions based on written schedule; Q1 asks about basic information, Q2 is based on an incorrect assumption or requires the test taker to make an inference, Q3 is a summary of multiple pieces of information	Listening stimulus: voice mail message that represents a problem or issue that requires the test taker to summarize and propose a solution (120–135 words)	Listening stimulus: prompt that is both seen and heard and requires test taker to take stance on an issue or topic
Prep time	45 sec.	30 sec.	0 sec.	0 sec.	30 sec.	15 sec.
Response time	45 sec	45 sec	15, 15, 30 sec.	15, 15, 30 sec.	60 sec.	60 sec.
Total time	Approximately 30 minutes for 11 questions					

Scoring Guidelines for the TOEIC Speaking and Writing Tests

The TOEIC Speaking Test Scoring Guidelines

During the TDA discussions and the development of the task shells, assessment specialists revisited several different kinds of scoring guidelines including those used for the TSE test and the TOEFL iBT Speaking test. Ultimately, it was decided that the TOEFL iBT Speaking test rubrics would provide the underlying foundation for the most difficult tasks on the TOEIC Speaking test. However, the design team agreed upon a five-band scoring scale rather than staying with the TOEFL Speaking test's four-band scale because the population taking the TOEIC tests differ in that it represents more test takers in the lower range of ability. The inter-rater reliability results for a five-band scoring scale were comparable to inter-rater reliability statistics for the TOEFL Speaking test's four-band scale. Considering the nature of the intermediate- and lower-level tasks, a scoring scale with fewer levels seemed to best represent the less-demanding requirements for these easier tasks. Adopting a three-band scale for these nine was more appropriate for these levels of tasks and also had the benefit of facilitating more accurate and efficient scoring.

Raters were directed to use task appropriateness as the driving consideration in assigning holistic scores because the test is task-based. Characteristics of responses are described for each of the levels represented in the scoring guides (Appendices A–E). The following response characteristics were developed beginning with the areas of measurement specified in the task shell with the exception of the analytically scored read a text aloud tasks: task appropriateness; linguistic resources (especially range and complexity of structures and vocabulary; clarity and pace of speech); discourse competence (especially cohesion and progression of ideas in sustained speech); and content (especially relevance, accuracy and completeness). The details for each level description were filled in based on the analysis of the tasks piloted in the field study. For the field study, there were originally six different scoring rubrics representing each of the six different task types on the speaking test.

In May of 2008, after eighteen months of operational testing, enough data had been collected to justify revising some of the language in the scoring guidelines to make the language more accessible to raters; therefore, making scores easier to assign. Two notable revisions were made:

1. The language of the scoring guides for questions 4–6 and questions 7–9 were combined into one scoring guide. The language overlapped enough that collapsing these rubrics into one provided an opportunity to make the scoring process more efficient and decreased the burden on raters to manage multiple scoring guides.

2. The language for each of the hardest tasks (responding to a voice mail message by proposing a solution and expressing an opinion) needed to be refined. As test takers were becoming more savvy at implementing test preparation strategies, it was evident that more and more responses, specifically to the propose a solution task, were beginning to sound relatively good, but were so generic that the response could be one given to any question in any test form (i.e. there was no language to connect the response to the specific situation presented in the stimulus). The description of this category of generic response was refined in the scoring guides to acknowledge an increasingly common trend in responses so that raters were directed to the most appropriate score.

For the field study, 20 TOEFL Speaking test raters scored 2,537 candidates' tests across four different test forms. A portion of the responses were double-scored. The analyses of this data confirmed the hypothesis that candidates who performed well on the Claim 3 tasks also performed well on the Claim 2 and Claim 1 tasks.

Writing - The results of the TDA for the writing measure are summarized in Table 8. The general claim for the writing measure, that test takers can use written English to perform typical international workplace communication tasks, remained unchanged after the field study. Before conducting the field study, the test designers thought that they would be able to predict which kinds of words would be easier or harder to implement into a grammatically correct sentence in questions 1–5, so all of the tasks were tagged for difficulty — easy, medium and difficult. The field study results were mixed and did not confirm the hypothesis that difficulty could be controlled at this level of detail for this simple task. Based on analyses of the field study results, the revised task specifications were revised to group item difficulty into two broader categories of easier and harder word combinations.

Like speaking, three subclaims were identified for writing that provided support for the overall claim that a test taker can use written English to perform typical international workplace communication tasks. Again, these subclaims were ordered hierarchically with the assumption that those who perform well on the hardest level task will also perform well on the intermediate- and beginning-level tasks. Task types that would provide evidence related to each of the three subclaims were defined and can be seen in Table 8. Because one of the business requirements outlined early in the process was to have the possibility of administering the TOEIC Writing test in a paper-based format, no audio components were included in the test design in order to more easily facilitate this mode of delivery. The stimulus materials for the writing tasks represented appropriate contexts and language for the claim for which they were providing evidence.

TABLE 8

Summary of Specifications for Writing Measure of the TOEIC Test

Writing claim	The test taker can use written English to perform typical international workplace communication tasks		
Subclaims	The test taker can produce well-formed sentences (including subordination).	The test taker can produce multi-sentence-length text to convey straightforward information, questions, instructions, narratives, etc.	The test taker can produce multi-paragraph-length text to express complex ideas, using, as appropriate, reasons, evidence and extended explanations.
Nature of writing task	Write a sentence based on a picture	Respond to a written request in the form of an e-mail	State, explain and support an opinion on an issue
Scoring rubric	Independent 0–3	Independent 0–4	Independent 0–5
Number of Questions	5	2	1
Nature of stimulus material	The photograph can be in color or black and white. Features of the words below the picture: 3. Two key words 4. Key words appear in all lowercase letters. 5. Key words appear separated by a space, then a slash “/”, then another space. Each form will have a balance of easy and difficult word combinations	The request will appear in e-mail format • 25–50 words long • Presents a situation to which the test taker must respond • The stimulus should not ask questions that the required tasks do not also ask.	The prompt will be a maximum 50 words in length and presents an ethical, moral or practical tension. The prompt asks for support of an opinion and is accessible to an international audience.
Response time	8 minutes	10 minutes for each question	30 minutes
Total time	Approximately 1 hour for 8 questions		

Since the TDA discussions and the development of the task shells were being conducted by the same assessment specialists who worked on the speaking test, the process for identifying appropriate scoring criteria was similar. The test designers revisited several different kinds of scoring rubrics including those used for the TWE® test and the TOEFL iBT Writing test. The most difficult task on the writing test was identical to the independent task for the TOEFL iBT Writing test. Since the scoring guides, though evaluating a different test population, had been used successfully for more than 20 years in the TOEFL test program, the test designers found it appropriate to use the same scoring criteria for this difficult task. Considering the nature of the intermediate- and lower-level tasks, creating a four-band scale for the intermediate tasks and a three-band scale for the beginning tasks were more appropriate for these levels of tasks and also had the benefit of facilitating more accurate and efficient scoring.

Characteristics of responses are described for each of the levels represented in the scoring guides (Appendices F-H). These response characteristics were developed beginning with the areas of measurement specified in the task shell: task appropriateness, organization, coherence, and fluency in language use (especially syntactic variety, appropriate word choice). The details for each level were filled in based on the analysis of the piloted sample tasks.

For the field study, 15 TOEFL Writing test raters scored 2,537 candidates' tests across four different test forms. A portion of the responses were double-scored. The analyses of this data confirmed the hypothesis that the test designers had made that candidates who performed well on the Claim 3 tasks also performed well on the Claim 2 and Claim 1 tasks (see Chapter 10).

Conclusion

As a result of the ECD process described in this chapter, modifications were made to tasks to solve the problems that had appeared during the field study. The TDA process required careful thinking and further specification of all of the tasks in the draft blueprint. The outcome from this process was the detailed task specifications required for developing and administering the operational test. The field study of the test blueprint established the psychometric properties of the TOEIC Speaking and Writing tests. The next steps were to provide test users with information about the new test.

References

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph No. MS-20). Princeton, NJ: ETS.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph No. MS-18). Princeton, NJ: ETS.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph No. MS-16). Princeton, NJ: ETS.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19, 477–496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. Measurement: *Interdisciplinary Research and Perspectives*, 1, 3–62.

Appendix A

TOEIC Speaking Test Read a Text Aloud Analytic Scoring Guides

QUESTIONS 1 AND 2

Pronunciation

Score	Response Description
3	Pronunciation is highly intelligible, though the response may include minor lapses and/or other language influence.
2	Pronunciation is generally intelligible, though it includes some lapses and/or other language influence.
1	Pronunciation may be intelligible at times, but significant other language influence interferes with appropriate delivery of the text.
0	No response OR no English in the response OR response is completely unrelated to the test.

Intonation & Stress

Score	Response Description
3	Use of emphases, pauses, and rising and falling pitch is appropriate to the text.
2	Use of emphases, pauses, and rising and falling pitch is generally appropriate to the text, though the response includes some lapses and/or moderate other language influence.
1	Use of emphases, pauses, and rising and falling pitch is not appropriate, and the response includes significant other language influence.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Appendix B

TOEIC Speaking Test Describe a Picture Scoring Guide

QUESTION 3

Score	Response Description
3	The response describes the main features of the picture. • The delivery may require some listener effort, but it is generally intelligible. • The choice of vocabulary and use of structures allows coherent expression of ideas.
2	The response is connected to the picture, but meaning may be obscured in places. • The delivery requires some listener effort. • The choice of vocabulary and use of structures may be limited and may interfere with overall comprehensibility.
1	The response may be connected to the picture, but the speaker's ability to produce intelligible language is severely limited. • The delivery may require significant listener effort. • The choice of vocabulary and use of structures is severely limited OR significantly interferes with comprehensibility.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Appendix C

TOEIC Speaking Test Respond to Short Questions Scoring Guide

QUESTIONS 4–9

Score	Response Description
3	The response is a full, relevant, socially appropriate reply to the question. In the case of Questions 7–9, information from the prompt is accurate. • The delivery requires little listener effort. • The choice of vocabulary is appropriate. • The use of structures fulfills the demands of the task.
2	The response is a partially effective reply to the question, but is not complete, fully appropriate, or in the case of Questions 7–9, fully accurate. • The delivery may require some listener effort, but is mostly intelligible. • The choice of vocabulary may be limited or somewhat inexact, although overall meaning is clear. • The use of structures may require some listener effort for interpretation. • In the case of Questions 7–9, the speaker may locate the relevant information in the prompt but fail to distinguish it from irrelevant information or fail to transform the written language so a listener can easily understand it.
1	The response does not answer the question effectively. Relevant information is not conveyed successfully. • The delivery may impede or prevent listener comprehension. • The choice of vocabulary may be inaccurate or rely on repetition of the prompt. • The use of structures may interfere with comprehensibility.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Appendix D

TOEIC Speaking Test Propose a Solution Scoring Guide

QUESTION 10

Score	Response description
5	The response successfully completes all parts of the task and is readily intelligible, coherent and sustained. It is characterized by ALL of the following: • The speaker plays the appropriate role and understands the relationship between the sender and the receiver of the message. • A clear understanding of the situation in the prompt and a relevant, detailed response to the situation is present. • The speech is clear with a generally well-paced flow. • Good control of basic and complex structures, as appropriate, is exhibited. Some minor errors may be noticeable, but they do not obscure meaning. • The use of vocabulary is effective, with allowance for minor inaccuracy.
4	The response addresses all parts of the task appropriately, but may fall short of being fully developed. It is generally intelligible, sustained and coherent, with some minor lapses. • The speaker plays the appropriate role and understands the relationship between the sender and the receiver of the message. • The response is sustained and conveys the minimum relevant information required by the situation in the prompt. • Minor difficulties with pronunciation, intonation or pacing are noticeable and may require listener effort at times although overall intelligibility is not significantly affected. • The response demonstrates fairly automatic and effective use of grammar but may be somewhat limited in the range of structures used. • The use of vocabulary is fairly effective. Some vocabulary may be inaccurate or imprecise.
3	The response attempts to address the task, but does not successfully complete all parts of the task. It contains mostly intelligible speech, although problems with delivery and/or overall coherence may occur. • The speaker may neglect the role-playing aspect of the task or misrepresent the relationship between the sender and the receiver of the message. • The response conveys some relevant information, but is clearly incomplete or inaccurate or the response is based on a misunderstanding of the task or content of the stimulus. • The speech is basically intelligible, although listener effort may be needed because of unclear articulation, awkward intonation or choppy rhythm/pace. • The response demonstrates limited control of grammar. • The use of vocabulary is limited.

Score	Response description
2	The response includes very little relevant content and/or speech is mostly unintelligible or incoherent. The content may be limited because of the following: • There are lengthy, socially inappropriate pauses. • The response is only tangentially related to the stimulus and tasks. The speech may be mostly unintelligible because of the following: • The delivery is labored and requires considerable listener effort. • There is very limited control of grammar. • The use of vocabulary is severely limited or inexact.
1	The response may be completely unintelligible OR the response may consist of isolated words or phrases, or mixtures of the first language and English OR the response may be vague and general and show no interaction with the prompt.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Appendix E

TOEIC Speaking Test Express an Opinion Scoring Guide

QUESTION 11

Score	Response Description
5	The response clearly indicates the speaker’s choice or opinion, and support of the choice or opinion is readily intelligible, sustained, and coherent. It is characterized by ALL of the following: • The speaker’s choice or opinion is supported with reason(s), details, arguments or exemplifications; relationships between ideas are clear. • The speech is clear with generally well-paced flow. It may include minor lapses or minor difficulties with pronunciation or intonation patterns that do not affect overall intelligibility. • Good control of basic and complex structures, as appropriate, is exhibited. Some minor errors may be noticeable, but they do not obscure meaning. • The use of vocabulary is effective, with allowance for occasional minor inaccuracy.
4	Response clearly indicates the speaker’s choice or opinion and adequately supports or develops the choice or opinion. • The response explains the reason(s) for the speaker’s choice or opinion, although the explanation may not be fully developed; relationships between ideas are mostly clear, with occasional lapses. • Minor difficulties with pronunciation, intonation or pacing are noticeable and may require listener effort at times, although overall intelligibility is not significantly affected. • The response demonstrates fairly automatic and effective use of grammar, but may be somewhat limited in the range of structures used. • The use of vocabulary is fairly effective. Some vocabulary may be inaccurate or imprecise.
3	The response expresses a choice, preference or opinion, but development and support of the choice or opinion is limited. • The response provides at least one reason supporting the choice, preference, or opinion. However, it provides little or no elaboration of the reason, repeats itself with no new information, is vague or is unclear. • The speech is basically intelligible, though listener effort may be needed because of unclear articulation, awkward intonation or choppy rhythm/pace; meaning may be obscured in places. • The response demonstrates limited control of grammar; for the most part, only basic sentence structures are used successfully. • The use of vocabulary is limited.

Score	Response Description
2	The response states a choice, preference or opinion relevant to the prompt, but support for the choice, preference or opinion is missing, unintelligible or incoherent. • Consistent difficulties with pronunciation, stress and intonation cause considerable listener effort; delivery is choppy, fragmented or telegraphic; there may be long pauses and frequent hesitations. • Control of grammar severely limits expression of ideas and clarity of connections among ideas. • The use of vocabulary is severely limited or highly repetitious.
1	The response is limited to reading the prompt or the directions aloud OR the response fails to state an intelligible choice, preference or opinion as required by the prompt OR the response consists of isolated words or phrases or mixtures of the first language and English.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Appendix F

TOEIC Writing Test Write a Sentence Based on a Picture Scoring Guide

QUESTIONS 1–5

Score	Response Description
3	The response consists of ONE sentence that: • has no grammatical errors, • contains forms of both key words used appropriately, AND • is consistent with the picture.
2	The response consists of one or more sentences that: • have one or more grammatical errors that do not obscure the meaning, • contain BOTH key words, (but they may not be in the same sentence and the form of the word(s) may not be accurate), AND • are consistent with the picture.
1	The response: • has errors that interfere with meaning, • omits one or both key words, OR • is not consistent with the picture.
0	The response is blank, written in a foreign language, or consists of keystroke characters.

Appendix G

TOEIC Writing Test Respond to a Written Request Scoring Guide

QUESTIONS 6 AND 7

Score	Response Description
4	The response effectively addresses all the tasks in the prompt using multiple sentences that clearly convey the information, instructions, questions, etc., required by the prompt. • The writer uses organizational logic or appropriate connecting words or both to create coherence among sentences. • The tone and register of the response is appropriate for the intended audience. • A few isolated errors in grammar or usage may be present, but they do not obscure the writer's meaning.
3	The response is mostly successful but falls short in addressing one of the tasks required by the prompt. • The writer omits, responds unsuccessfully or responds incompletely to ONE of the required tasks. • The writer uses organizational logic or appropriate connecting words in at least part of the response. • The writer shows some awareness of audience. • Noticeable errors in grammar and usage may be present; ONE sentence may contain errors that obscure meaning.
2	The response is marked by several weaknesses: • The writer addresses only ONE of the required tasks or unsuccessfully or incompletely addresses TWO OR THREE of the required tasks. • Connections between ideas may be missing or obscure • The writer may show little awareness of audience • Errors in grammar and usage may obscure meaning in MORE THAN ONE sentence

Score	Response Description
1	The response e-mail is seriously flawed and conveys little or no information, instructions, questions, etc., required by the prompt. • The writer addresses NONE of the required tasks, although the response may include some content relevant to stimulus. • Connections between ideas are missing or obscure. • The tone or register may be inappropriate for the audience. • Frequent errors in grammar and usage obscure the writer's meaning most of the time.
0	A response at this level merely copies words from the prompt or stimulus, rejects the topic or is otherwise not connected to the topic, is written in a language other than English, consists of keystroke characters that convey no meaning, or is blank.

Appendix H

TOEIC Writing Test Write an Opinion Essay Scoring Guide

QUESTION 8

Score	Response Description
5	A response at this level largely accomplishes all of the following: • It effectively addresses the topic and task. • It is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details. • It displays unity, progression and coherence. • It displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors.
4	A response at this level largely accomplishes all of the following: • It addresses the topic and task well, though some points may not be fully elaborated. • It is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details. • It displays unity, progression and coherence, though it may contain occasional redundancy, digression or unclear connections. • It displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning.
3	A response at this level is marked by one or more of the following: • It addresses the topic and task using somewhat developed explanations, exemplifications, and/or details. • It displays unity, progression and coherence, though connection of ideas may be occasionally obscured. • It may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning. • It may display accurate but limited range of syntactic structures and vocabulary.
2	A response at this level may reveal one or more of the following weaknesses: • Limited development in response to the topic and task • Inadequate organization or connection of ideas • Inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • A noticeably inappropriate choice of words or word forms • An accumulation of errors in sentence structure and/or usage

Score	Response Description
1	A response at this level is seriously flawed by one or more of the following weaknesses: • Serious disorganization or underdevelopment • Little or no detail, irrelevant specifics, or questionable responsiveness to the task • Serious and frequent errors in sentence structure or usage
0	An essay at this level merely copies words from the prompt, rejects the topic or is otherwise not connected to the topic, is written in a language other than English, consists of keystroke characters that convey no meaning, or is blank.