ETS®

TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

TOEIC® Listening and Reading Test Scale Anchoring Study

Chi-wen Liao

January 2010

Over the past few years, the TOEIC® test examinees and test users have increasingly expressed the need to better understand the meaning of reported scaled scores, which are numeric scores and provide information pertaining only to an individual's relative performance on a specific TOEIC test scale. In addition to scaled scores, the TOEIC test customers would also like to obtain diagnostic-like information pertaining to a scaled score, such as what English–language skills an individual has and how well-developed these skills are. This information can play an instrumental role in activities related to learning, instruction and employment selection for both examinees and test users. To meet this customer need, a TOEIC test scale anchoring study was conducted at ETS in May 2005 as a part of the test development plan for the Redesigned TOEIC Listening and Reading test. The purpose of the study was to provide examinees with descriptive score proficiency information related to their TOEIC test scaled scores.

The study used the scale anchoring method, which is a method used by several well-known large-scale assessments such as the National Assessment of Educational Progress (NAEP; Beaton & Allen, 1992), the Trends in International Mathematics and Science Study (TIMSS; Kelly, 1999), and the TOEFL iBT™ test (TOEFL iBT™; Zhang, 2006). This paper describes: (a) the data and the procedures used to develop the proficiency descriptions, (b) how to use and interpret the score proficiency information, (c) how the score proficiency descriptions were added in the redesigned TOEIC test score report, and (d) limitations of the study and future plans.

Test

This study was conducted using data from examinees (N = 1,958) who participated in the field study for the Redesigned TOEIC Listening and Reading test because operational data for the redesigned test were not available at that time. The field study was conducted in November 2004 (see Chapter 3 in this compendium). The examinees, who were from the TOEIC test Secured Program (SP) population in Japan and Korea, took two field test forms, forms C and D. The field study sample was found to have higher listening scores and slightly lower reading scores than the population of the Redesigned TOEIC Listening and Reading test from May 2006 to June 2009. For the field study sample, the means (standard deviations) were 341 (80) and 270 (90) for listening and reading skills, respectively. For the Redesigned TOEIC test population, the means (standard deviations) were 323 (85) and 278 (95) for listening and reading skills, respectively. The field study sample was considered to be reasonably close to the population because the standardized difference between the group means was .22 for the listening section of the test and .09 for the reading section.

Like the Classic TOEIC Listening and Reading test, the redesigned test has 200 items, with 100 items in each section. Like the Classic TOEIC Listening and Reading test, the score scale for each section of the redesigned test is still 5 to 495 points, in increments of 5 points. Although the listening and reading sections both have the same scale, their scaled scores cannot be compared to each other as they were derived from independent separate equatings.

Scale Anchoring Method

One way in which the TOEIC test program determines the meaning of TOEIC test scores is by examining how examinees have performed on the questions that comprise the test. More specifically, the TOEIC test program determines which questions examinees can answer correctly at various score levels. Content specialists then examine these questions and ascertain what the questions have in common. On this basis, the content specialists make inferences about what examinees at different score levels are able to do in terms of English language skills. The process is generally called *scale anchoring*, and the technical aspects are described in detail below.

The scale anchoring method was first proposed by Beaton and Allen (1992) for NAEP. A scale anchoring usually includes the following steps:

1. Specify the score proficiency levels to be described.
2. Calculate the conditional percentage of examinees answering an item correctly (conditional p-value) at each of the specified proficiency levels for all the items in an item pool.
3. Identify anchor items that describe examinees' performance at each proficiency level.
4. Identify the skills and knowledge measured by the anchor items, using the expert judgments of content specialists.

Details of how these steps were carried out in this study are described in the text that follows.

Step 1: Specify Score Proficiency Levels

Several sets of proficiency score levels were tentatively chosen for the listening and reading sections. These tentative levels were then analyzed following Steps 2 and 3. The goal was to select a set of score levels for each measure that could meet two requirements: (a) there should be a sufficient number of levels to adequately define the scale, and (b) the proficiency levels should be separated widely enough to include a sufficient number of anchor items. Because these two requirements often are in conflict, efforts were made to find a balance between the two requirements based on available data.

After several trials, the proficiency levels were selected and finalized by statisticians and content specialists at the scaled scores of 200, 300, 400 and 460 for the listening section, and 150, 250, 350 and 450 for the reading section, based on the data from two redesigned TOEIC field test forms.

Step 2: Calculate Conditional P-Value

Examinees in the field test who scored about one standard error of measurement (SEM) above and below each proficiency score level were grouped together. The SEM for each of the listening and reading sections is about 25 scaled score points. Altogether there were four listening groups and four reading groups. The conditional item *p*-values were estimated by calculating the individual item *p*-values for all the items in the two field study forms for each of the groups separately.

Step 3: Identify Anchor Items Based on Two Discriminating Criteria

Based on the conditional *p*-values estimated in Step 2, an item was identified as an anchor item for a particular proficiency level if it met the following two criteria simultaneously:

1. The conditional p-value of the item at a proficiency level had to be .5 or higher, and the conditional p-value of the item at the adjacent next lower level had to be less than .5.
2. The difference between this item's conditional p-values at this proficiency level and the adjacent next lower level had to be .2 or higher.

An item that met one criterion but not both was called an undefined item. As an example, four items were selected as listening anchor items, and their associated conditional *p*-values across the four proficiency levels are given in Table 1.

TABLE 1

Four Listening Anchor Items and Their Conditional P-Values

Item #	All examinees	Level 200 examinees	Level 300 examinees	Level 400 examinees	Level 460 examinees
16	.83	.54[a]	.81	.94	.98
56	.70	.31	.62[a]	.89	.93
55	.45	.21	.31	.58[a]	.76
50	.34	.07	.19	.43	.72[a]

[a] The level at which the item was anchored.

Table 2 below summarizes the number of anchor items that were finally selected from each of the field test forms, forms C and D.

TABLE 2

Number of Anchor Items Identified From Forms C and D

	Listening			Reading	
Level	Form C	Form D	Level	Form C	Form D
200	18	21	150	25	27
300	20	25	250	18	20
400	24	21	350	24	14
460	10	7	450	13	12
Undefined	28	26	Undefined	20	27

Step 4: Identify Skills and Knowledge Measured by the Anchor Items

Content specialists examined the anchor items from Step 3 and determined what skills and knowledge were needed to provide a correct response. Using this information as a foundation, the content specialists developed a concise description of the types of English tasks that examinees at each proficiency level can and cannot do. These descriptions appear as statements of strengths and weaknesses in the TOEIC Listening and Reading test score proficiency tables (see Appendix A). The final proficiency tables do not include Level 460 for the listening section because after reviewing those anchor items at that level, the content specialists believed that the number of items was insufficient to allow them to reliably describe the characteristics of examinees at that level.

How to Use and Interpret the Score Proficiency Description Information

Although an item can be anchored at a particular level using the method described above, this does not mean that all examinees at that proficiency level can answer the question correctly. Rather, it means that the examinees, as a group, at that proficiency level, generally have a much higher probability of answering the item correctly than do examinees at the adjacent next lower proficiency level. For example, with Item 56 in Table 1, examinees scoring around 300 on the listening section collectively have a probability of .62 of answering the item correctly, but those scoring around 200 have only a probability of .31 of answering the item correctly. Thus, when interpreting the statements of strengths and weaknesses, one should keep in mind that the statements are group-level descriptions based on empirical data and human judgments, and they may not apply exactly to any particular individual.

The statements of strengths and weaknesses in the appendix are most appropriate for examinees scoring within one SEM of the specified levels. For those who score between two proficiency levels, for example, with a listening score between 240 and 270, which proficiency level statement is appropriate? The answer is that the statements for both the 200 and 300 levels should be referred to because these examinees are likely to possess a mix of characteristics typical of both levels. The closer an examinee’s score is to a defined level, the more likely the examinee possesses the characteristics defined for that level. That is, examinees scoring at 240 are likely to find more statements at the 200 level that are applicable to them than at the 300 level. On the other hand, examinees scoring at 270 are likely to find more statements at the 300 level that are applicable to them than at the 200 level.

Examinees scoring below the range of the 200 level on the listening section and the 150 level on the reading section may have some of the same strengths as test examinees scoring around 200 and 150, but their performance is likely to be less consistent than those who score right around the levels.

Adding Score Proficiency Information to the Redesigned TOEIC Test Score Report

One of the new features in the redesigned TOEIC test score report is the addition of score proficiency descriptions. Because of the limited space on the score report, only the statements of strengths for a particular level are printed. Examinees can refer to the TOEIC test official website at http://www.ets.org/toeic for the complete proficiency description tables.

The statements of strength for a particular score level are provided to examinees who score within one standard error of measurement (i.e., 25 scaled score points) of a score level. For example, any examinees scoring between 275 and 325 will be provided with statements related to score level 300. Examinees scoring between two levels are provided with the statements for only the lower level, but are advised to read statements for both levels from the complete tables. Examinees scoring below Level 200 on listening section or level 150 on the reading section are provided with the statements for these two levels.

Discussion

Based on the field study forms, this study found that it was not possible to statistically specify more than the current three levels for the listening section (200, 300 and 400) and four levels for the reading section (150, 250, 350 and 450) and at the same time to identify enough anchor items to discriminate between examinees at adjacent levels. The more levels that were specified, the fewer the number of items that could be anchored at a level. While this study identified as many anchor items as possible, the distance between levels had to be expanded. Thus, the final levels are some distance apart and examinees scoring between levels are instructed to review statements for both the level below and the level above. In July and August of 2006, ETS performed a statistical analysis based on several redesigned TOEIC test operational forms. The analysis examined whether the current number of levels could be expanded from three and four levels (for the listening and reading sections, respectively) to seven levels for each test section, as some test users had requested. The results indicated that such an increase was not possible at that time. In the future, the TOEIC test program plans to continue the effort to expand the number of levels, as statistically allowed, to meet the need of users. The program will also examine whether the descriptors for levels are applicable for examinees over time.

References

Beaton, A., & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics,* 17(2), 191-204.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring.* Retrieved October 21, 2009, from the Boston College Dissertations and Theses Web site: http://escholarship.bc.edu/dissertations/AAI9923420/

Zhang, Y. (2006). *Estimating the proficiency levels of TOEFL iBT Reading and Listening using scale anchoring method.* Unpublished manuscript.

Appendix A

TOEIC TEST LISTENING SCORE PROFICIENCY TABLE

Level	Strengths	Weaknesses
400	Test takers who score around 400 *typically* have the following strengths: • They can infer the central idea, purpose and basic context of *short* spoken exchanges across a broad range of vocabulary, even when conversational responses are indirect or not easy to predict. • They can infer the central idea, purpose and basic context of *extended* spoken texts across a broad range of vocabulary. They can do this even when the information is not supported by repetition or paraphrase and when it is necessary to connect information across the text. • They can understand details in *short* spoken exchanges, even when negative constructions are present, when the language is syntactically complex, or when difficult vocabulary is used. • They can understand details in *extended* spoken texts, even when it is necessary to connect information across the text and when this information is not supported by repetition. They can understand details when the information is paraphrased or when negative constructions are present.	Test takers who receive a score at this level *typically* have weaknesses only when uncommon grammar or vocabulary is used.

Level	Strengths	Weaknesses
300	Test takers who score around 300 *typically* have the following strengths: • They can sometimes infer the central idea, purpose and basic context of *short* spoken exchanges, especially when the vocabulary is not difficult. • They can understand the central idea, purpose and basic context of *extended* spoken texts when this information is supported by repetition or paraphrase. • They can understand details in *short* spoken exchanges when easy or medium-level vocabulary is used. • They can understand details in *extended* spoken texts when the information is supported by repetition and when the requested information comes at the beginning or end of the spoken text. They can understand details when the information is slightly paraphrased.	Test takers who score around 300 *typically* have the following weaknesses: • They have difficulty understanding the central idea, purpose and basic context of *short* spoken exchanges when conversational responses are indirect or difficult to predict or when the vocabulary is difficult. • They do not understand the central idea, purpose and basic context of *extended* spoken texts when it is necessary to connect information within the text or when difficult vocabulary is used. • They do not understand details in *short* spoken exchanges when language is syntactically complex or when difficult vocabulary is used. They do not usually understand details that include negative constructions. • They do not understand details in *extended* spoken texts when it is necessary to connect information across the text or when the information is not supported by repetition. They do not understand most paraphrased information or difficult grammatical constructions.
200	Test takers who score around 200 *typically* have the following strengths: • They can understand *short* (single-sentence) descriptions of the central idea of a photograph. • They can sometimes understand the central idea, purpose and basic context of *extended* spoken texts when this information is supported by a lot of repetition and easy vocabulary. • They can understand details in *short* spoken exchanges and descriptions of photographs when the vocabulary is easy and when there is only a small amount of text that must be understood. • They can understand details in *extended* spoken texts when the requested information comes at the beginning or end of the text and when it matches the words in the spoken text.	Test takers who score around 200 *typically* have the following weaknesses: • They do not understand the central idea, purpose or basic context of *short* spoken exchanges, even when the language is direct and no unexpected information is present. • They do not understand the central idea, purpose and basic context of *extended* spoken texts when it is necessary to connect information across the text or when the vocabulary is somewhat difficult. • They do not understand details in *short* spoken exchanges when somewhat difficult vocabulary is used or when the language is syntactically complex. They do not understand details that include negative constructions. • They do not understand details in *extended* spoken texts when the requested information is heard in the middle of the text. They do not understand paraphrased information or difficult grammatical constructions.

TOEIC TEST READING SCORE PROFICIENCY TABLE

Level	Strengths	Weaknesses
450	**Your scaled score is close to 450. Test takers who score around 450 *typically* have the following strengths:** • They can infer the central idea and purpose of a written text, and they can make inferences about details. • They can read for meaning. They can understand factual information, even when it is paraphrased. • They can connect information across an entire text, and they can make connections between two related texts. • They can understand a broad range of vocabulary, unusual meanings of common words and idiomatic usage. They can also make distinctions between the meanings of closely related words. • They can understand rule-based grammatical structures. They can also understand difficult, complex and uncommon grammatical constructions.	Test takers who score around 450 *typically* have weaknesses only when the information tested is particularly dense or involves difficult vocabulary.
350	Test takers who score around 350 *typically* have the following strengths: • They can infer the central idea and purpose of a written text, and they can make inferences about details. • They can read for meaning. They can understand factual information, even when it is paraphrased. • They can connect information across a small area within a text, even when the vocabulary and grammar of the text are difficult. • They can understand medium-level vocabulary. They can sometimes understand difficult vocabulary in context, unusual meanings of common words and idiomatic usage. • They can understand rule-based grammatical structures. They can also understand difficult, complex and uncommon grammatical constructions.	Test takers who score around 350 *typically* have the following weaknesses: • They do not connect information across a wide area within a text. • They do not consistently understand difficult vocabulary, unusual meanings of common words or idiomatic usage. They usually cannot make distinctions between the meanings of closely related words.

Level	Strengths	Weaknesses
250	Test takers who score around 250 *typically* have the following strengths: • They can make simple inferences based on a limited amount of text. • They can locate the correct answer to a factual question when the language of the text matches the information that is required. They can sometimes answer a factual question when the answer is a simple paraphrase of the information in the text. • They can sometimes connect information within one or two sentences. • They can understand easy vocabulary, and they can sometimes understand medium-level vocabulary. • They can understand common, rule-based grammatical structures. They can make correct grammatical choices, even when other features of language, such as difficult vocabulary or the need to connect information, are present.	Test takers who score around 250 *typically* have the following weaknesses: • They do not understand inferences that require paraphrase or connecting information. • They have a very limited ability to understand factual information expressed as a paraphrase using difficult vocabulary. They often depend on finding words and phrases in the text that match the same words and phrases in the question. • They usually do not connect information beyond two sentences. • They do not understand difficult vocabulary, unusual meanings of common words or idiomatic usage. They usually cannot make distinctions between the meanings of closely related words. • They do not understand more-difficult, complex or uncommon grammatical constructions.
150	Test takers who score around 150 *typically* have the following strengths: • They can locate the correct answer to a factual question when not very much reading is necessary and when the language of the text matches the information that is required. • They can understand easy vocabulary and common phrases. • They can understand the most-common, rule-based grammatical structures when not very much reading is necessary.	Test takers who score around 150 *typically* have the following weaknesses: • They cannot make inferences about information in written texts. • They do not understand paraphrased factual information. They rely on matching words and phrases in the text to answer questions. • They are often unable to connect information even within a single sentence. • They understand only a limited range of vocabulary. • They do not understand even easy grammatical constructions when other language features, such as difficult vocabulary or the need to connect information, are also required.