PASAA Journal
Volume 66, October 2023, 212–225

PASAA

Idea-Sharing

Crafting Item Difficulty in TOEFL iBT Listening Tests

Alan Shaw [a]*

Former Assessment Specialist, Educational Testing Service, Princeton, New Jersey, USA

**Corresponding author:* tallshaw2m10@gmail.com

APA citation:	Shaw, A. (2023). Idea-sharing: Crafting item difficulty in TOEFL iBT listening tests [Special issue]. *PASAA*, *66*, 212–225.

1. Introduction

Item difficulty in TOEFL iBT Listening tests is the product of interactions between two sets of complex relationships: 1) relationships among numerous item characteristics themselves, and 2) relationships between item characteristics and an extremely diverse international test taker population. As a former TOEFL test developer of 18 years' experience, in this article, I share my thoughts on the art of crafting item difficulty in TOEFL iBT Listening tests. My perspective incorporates elements of the training I received at ETS, and it also draws upon my years of experience collaborating with ETS colleagues and studying item performance statistics. It should be noted, however, that this paper should not be interpreted as a statement of ETS test development principles and no confidential ETS information is revealed in this paper.

2. The TOEFL iBT Test Developer's Task

Although the TOEFL iBT Listening test is sometimes used for other purposes, it was designed primarily for use as a college entrance examination. Universities often set a "cut score" for such tests, a minimum score that candidates must obtain to be considered for admissions. A common cut score

E-ISSN: 2287-0024

would be roughly equivalent to a requirement that candidates score in the top one-third of examinees taking the test. Thus, to provide universities with the information they need, the test must accurately discriminate among test takers of varying levels of listening skill, with especially keen discrimination among test takers whose skills position them near the borderline between the bottom two-thirds and the top one-third of test takers.

3. IRT vs. Classical Item Analysis Statistics

After each administration of a TOEFL iBT Listening test, the ETS statistics unit produces two sets of statistics for use by test developers: Item Response Theory (IRT) statistics and classical item analysis statistics. Test developers use IRT statistics to assemble operational test forms, as they make it possible to assemble forms that a) discriminate well at every test taker ability level within the ability range of greatest interest and b) are comparable in difficulty to other TOEFL iBT Listening test forms. For evaluating and understanding item performance, however, test developers generally find classical item analysis statistics more useful. For that reason, in the few instances in which statistics are referred to in this paper, the references will be to classical item analysis statistics.

4. Setting an Item Difficulty Target

There is a close relationship between item difficulty and item discrimination, and if a TOEFL iBT test is to provide the information universities need, it is essential that the difficulty of test items fall within the range of difficulty that provides the greatest potential for achieving a high degree of discrimination in the test taker ability range of greatest interest. In classical item analysis statistics, the measure of an item's difficulty is the average item score, i.e., the percentage of test takers who answered the item correctly. For example, if 45% of the test takers answered an item correctly, .45 is the item difficulty. In practice, .67 (the boundary between the bottom two-thirds and top one-third of test takers) is a useful target. That target will not often be hit precisely, but items whose difficulties fall within

15 or so percentage points of .67 are compatible with IRT difficulty targets and have substantial potential for discriminating well at the test taker ability levels where discrimination counts most. Items whose difficulties fall below or above that range may also provide useful information, but that information is of somewhat less interest and discrimination tends to decrease as item difficulty becomes more distant from the heart of the target item difficulty range.

5. The Influence of Text Types on Approaches to Crafting Item Difficulty

The text types in TOEFL iBT Listening tests are 1) monologic lectures, 2) interactive lectures (in which students interact with the professor in the course of a lecture), 3) one-on-one professor/student conversations, 4) conversations between a student and an employee at the university (e.g., between a student and a member of the housing office staff), and 5) student/student conversations. A full range of item difficulties can be achieved in the context of any of these text types. Each text type, however, presents opportunities to craft item difficulty in a different way. For example, to craft item difficulty in a lecture, a test developer might lean toward testing lower-frequency academic topics and lower-frequency academic vocabulary; to craft item difficulty in a professor/student conversation, a test developer might test language related to requirements for a research paper; to craft item difficulty in a conversation between a student and a university employee, a test developer might test language related to university policies; to craft item difficulty in a student/student conversation, a test developer might, for example, test colloquial language related to the students' plans for an upcoming spring vacation..

6. Categories of Difficulty Drivers

Test developers achieve item difficulty targets by manipulating *difficulty drivers*, factors that play a role in determining the difficulty of a test item. Over the course of my career at ETS, I came to think of difficulty drivers in TOEFL iBT Listening tests as falling into three categories: 1) those related to the *text* (i.e., the

words) that will be recorded, 2) those related to the *recording* of that text, and 3) those related to the *test questions* that test takers will answer after listening to the recorded text.

7. Difficulty Drivers Related to the Text

7.1 Topic

The topic of a lecture or conversation is a powerful difficulty driver. In a test like TOEFL iBT, whose test taker population is extremely diverse, it is essential that test developers "level the playing field" by selecting, to the extent possible, topics that are accessible to the entire TOEFL test taker population, regardless of ethnicity, gender, socio-economic status, or field of study. No matter what topic a test developer chooses, however, some test takers will come to the topic with more related background knowledge than others. To minimize the advantages and disadvantages associated with test takers' background knowledge, test developers can craft into listening texts any topic-specific background information that is essential for understanding the topic of the text.

Example of crafting background information into a text:

(Professor in a history class, introducing a lecture on the first president of the United States) "*Okay, so, as we know, the country that would become known as the United States of America was born when British colonies in North America gained their independence from Great Britain in 1783."*

The history of the origin of the United States will be well known to test takers in some parts of the world, and completely unknown to test takers in other parts of the world. In the above example, the test developer has woven that information into the text in order to minimize the disadvantage of test takers who come to the test without that background knowledge.

While it is important that a topic be accessible to all test takers, it is equally important that a topic is not overly familiar to test takers. When a topic is too familiar, test takers can draw upon outside knowledge to answer test questions, so listening skills are not tested, and it is difficult to craft items that discriminate well between lower-ability and higher-ability test takers.

Example of an overly familiar topic:

(Professor in a marketing class) "Today we're going to talk about some of the strategies that supermarkets use to get shoppers to spend more money. First, a lot of supermarkets have a one-way entry door; you can use the door to get in the supermarket, but you can't use it to get out. To get out, you have to first walk through other parts of the supermarket, where you might see other things that you weren't planning to buy, but since they're right there in front of you, you buy them."

Conversely, when a lecture topic is outside the realm of most test takers' experience, test takers may not have the background they need in order to be able to efficiently process, upon a single hearing, the language they hear.

Example of an inappropriately unfamiliar topic:

(Professor in a biology class) "Toward the end of the last class, I recounted how researchers mixed various types of mouse embryonic stem cells and created artificial embryo cells, and how another team of researchers subsequently used a similar process to generate mouse embryoids that were very similar to genuine embryos."

In the author's experience, the topics that provide the best basis for measuring test takers' listening skills are those that present unfamiliar content in a familiar context. In the following lecture, for example, the principal elements—

meteorites, the Moon, and the Earth—will be familiar to all test takers. The concept of lunar meteorites will, however, be unfamiliar to most test takers.

Example of an appropriate topic:

(Professor in an astronomy class) "As we know, a meteorite's a chunk of rock that originates somewhere in outer space and falls to earth. Uh, and back in 1982, a scientist on an expedition in Antarctica found an unusual meteorite and sent it to a lab in Washington, DC, where another scientist examined it and noticed that it resembled some of the rocks that'd been brought back from the Moon by the Apollo space program. Further analysis indicated that this meteorite had, in fact, originated on the Moon, that it was what would become known as a lunar meteorite."

7.2 Vocabulary

Vocabulary is a powerful difficulty driver that can be used to craft subtle degrees of item difficulty. As would be expected, texts worded in higher-frequency vocabulary tend to yield easier items, texts worded in lower-frequency vocabulary tend to yield more difficult items.

Example of an idea expressed first in lower-frequency vocabulary, then in higher-frequency vocabulary:

Lower-frequency vocabulary: (Professor in a Thai language class) "The pitch and contour characteristics traditionally ascribed to any particular Thai tone are valid solely in the citation form of monosyllabic words spoken in isolation."

Higher-frequency vocabulary: (Professor in a Thai language class) "The only time you're going to hear the pitch and contour of a Thai tone pronounced the way

they're taught in school is when that tone's spoken with very proper pronunciation, in a one-syllable word, by itself, not in a sentence."

Although the information the professor presents in these two examples is essentially the same, items related to the higher-frequency-vocabulary version will likely be significantly easier.

8. Salience of Tested Information

Unlike in tests of reading comprehension, examinees taking tests of listening comprehension do not have an opportunity to review the text in order to locate answers to test questions. Therefore, it is essential that in listening tests all tested material be made salient enough that at least higher-ability test takers will recognize it as important and have sufficient opportunity to process it. Many difficulty drivers in listening tests are closely related to the degree of salience of tested information.

8.1 Repetition of Tested Information

Any form of repetition of tested information; verbatim repetition of the information, a gloss or paraphrase of the information, or a simple reference to the information, will make the information more salient and easier for the listener to process and remember, and will likely decrease the difficulty of the related item. Repetition of tested information is a powerful difficulty driver.

Example of repetition of tested information:
Test Question: What is the largest mammal in the world?

Compare:

Tested Text, Version 1: (Professor in a biology class) "Antarctic blue whales are the largest mammals in the world. They can weigh up to two hundred tons and reach up to thirty meters in length."

Tested Text, Version 2: (Professor in a biology class) "Antarctic blue whales are the largest mammals in the world. Antarctic blue whales can weigh up to two hundred tons and reach up to thirty meters in length."

The simple repetition of 'Antarctic blue whales" in the second text makes the tested information more salient than in the first text and will likely make the related item significantly easier.

8.2 Information Density

Closely related to repetition of tested language is information density. If new information is not subsequently repeated, paraphrased, or referred to, but is, rather, immediately followed by additional new information, the text may become overly information dense. In tests like TOEFL iBT Listening, test takers do not know as they listen what information will be tested, so they may try to remember everything they hear, so they may feel overwhelmed if too much information is presented too quickly. And even if some test takers are able to remember all of the information in an information-dense text, the text will likely be measuring their memory skill as much as it is testing their listening skill.

Example of excessive information density:

(Professor in a History of the English Language class) "The English language began with the migration of the Jutes, Angles, and Saxons from Germany and Denmark to Britain in the 5th and 6th centuries. The Norman Conquest of 1066 brought many French words into English, and Greek and Latin words began to enter English in the 15th century."

In a TOEFL iBT Listening test, most test takers would be overwhelmed by such a text. Even if, for example, only two of the facts presented in this text were subsequently tested, test takers would not know in advance which two facts those

would be. A good rule of thumb for avoiding excessive information density in listening texts is if the information is not going to be tested, it should not be included in the text. If the information is going to be tested, craft support for it by manipulating difficulty drivers as needed to make it salient and memorable.

8.3 Position of Tested Information in Its Sentence

A second difficulty driver related to the fleeting nature of a listening text is the position of tested information in its sentence. All else being equal, tested information positioned at the end of a sentence will be more salient than information positioned elsewhere in a sentence. This increased salience is related to the brief pause that typically follows the end of a sentence and gives test takers a brief moment to process what they have just heard before having to process new information.

Examples of position of tested information in sentence:

Test Question: In what year did Beethoven publish his first work?
Compare:

Tested Text, Version 1: (Professor in a music class) "Beethoven published his first work in 1783, under the tutelage of Christian Gottlob Neefe."

Tested Text, Version 2: (Professor in a music class) "Beethoven published his first work under the tutelage of Christian Gottlob Neefe, in 1783."

In Version 2, positioning the tested information, "in 1783," at the end of the sentence makes it more salient than in Version 1 and will likely decrease the difficulty of the related item.

9. Difficulty Drivers Related to the Recording of the Text

Three very powerful difficulty drivers are related to the recording of the text. These drivers can increase, decrease, or even completely negate the influence of the text-related difficulty drivers discussed in the previous section. It is therefore essential that recording sessions be directed by a test developer who understands how to achieve target item difficulty by balancing text-related and recording-related difficulty drivers.

Speech rate

One recording-related difficulty driver is the speaker's speech rate. As would be expected, the faster the speech rate, the more difficult the related items are likely to be; the slower the speech rate, the easier the related items are likely to be.

9.1 Degree of Emphasis the Speaker Places on Tested Information

A second recording-related difficulty driver is the degree of emphasis the speaker places on tested information. Again, as would be expected, the more emphasis the speaker puts on the tested information, the more salient the information, and the easier the related item is likely to be.

Example of the speaker's emphasis on tested information:

Test Question: What is the price of a new textbook for the student's class?
Compare:

Tested Text, Version 1: (student) "New, the textbook was almost eight thousand baht."

Tested Text, Version 2: (student) "New, the textbook for the class was almost EIGHT THOUSAND BAHT!"

The speaker's emphasis on the number 'eight thousand' makes this information more salient and more easily remembered, and will likely make the related item significantly easier.

9.2 Presence or Absence of a Pause After Tested Information

The third recording-related difficulty driver is the presence or absence of a pause after tested information. When tested information is followed by a pause, the test taker has a moment to process that information before new information is introduced, and the related item will likely be easier than if the tested information was immediately followed by additional information.

Example illustrating a pause following tested information:

Test Question: Where are the copy machines located?
Compare:

Tested Text, Version 1: (Library employee) "The copy machines are on the second floor, copies are three baht each."

Tested Text, Version 2: (Library Employee) "The copy machines are on the second floor. [pause] And copies are three baht each."
In Version 2, the pause after the first sentence makes the tested information, "on the second floor," more salient and will likely decrease the difficulty of the related item.

10. Difficulty Drivers Related to Test Questions

10.1 Item Type

The first difficulty driver related to the test questions is the item type. Gist items (which test examinees' skill in recognizing the general topic of a text) tend to be the easiest. Items testing explicit information tend to be more difficult than

gist items, and items testing implicit information tend to be more difficult than either gist items or items testing explicit information.

10.2 Attractiveness of Distracters

A second difficulty driver related to the questions is the attractiveness of the distracters, or the incorrect answer choices. For an item to discriminate well, two things need to happen: more highly-skilled test takers should be able to answer the item correctly and less highly-skilled test takers should be able to answer the item incorrectly. It is therefore essential that distracters a) are readily recognized as being incorrect by test takers who know the correct answer and b) are at the same time attractive to test takers who do not know the correct answer.

Three elements make a distracter attractive to less-highly-skilled test takers:

1) The distracter targets a predictable misinterpretation of the tested information:

Example:

Tested Text: [Professor in a history of cinema class] "The subject of the story told in the Academy Award-winning film Braveheart *was a Scotsman named William Wallace."*

Test Question: Who was William Wallace?

Distracter: The director of the movie Braveheart.

The mention of an Academy Award would likely make this distracter an attractive guess for many lower-ability test takers.

2) Language in the distracter matches salient language in the tested text:

Example:

Tested Text: [Professor in an art history class] "The subject of Leonardo da Vinci's painting Mona Lisa is believed to be Lisa Gherardini, the wife of a wealthy silk merchant."

Test Question: Who was Lisa Gherardini?

Distracter: a silk merchant

3) Language in the distracter matches salient language in the test question.

Example:

Test Question: Why did the ancient Greek scholar Archimedes shout "Eureka! Eureka!" when he stepped into a bathtub?

Distracter: The water in the bathtub was very hot.

CAUTION! One type of red flag in post-test item analyses is that a distracter was selected as the correct answer by a high percentage of higher-ability test takers. Upon carefully studying such items, test developers often realize that the distracter, when interpreted in an unexpected way, could, in fact, reasonably be considered a correct response. It is therefore important that test developers, when authoring test items, explore all possible interpretations of each distracter. (NOTE: When a flawed item such as this is discovered in post-test-administration

analyses, the item is dropped from scoring, and test takers' scores are adjusted to compensate.)

10.3 Ease of Processing the Test Questions and the Answer Options

Finally, a third difficulty driver related to the questions is the degree of ease with which a test question and its related answer options can be processed by test takers. Questions and answer options that are for any reason—excessive length, overly complex syntax, low frequency vocabulary, confusing wording—difficult to process can greatly increase item difficulty, for the wrong reason. Such questions and answer options can easily override the influence of all previously employed construct-relevant difficulty drivers and ruin the measurement of test-takers' listening skills. While achieving target item difficulty is important, test developers must keep in mind that item difficulty is meaningful *only if it is crafted in a way that provides a valid, fair, and accurate measure of test takers' English listening comprehension ability.*

11. About the Author

Alan Shaw holds a B.A degree in history from Duke University and an M.A. degree in TESOL from California State University Los Angeles. Alan worked as an assessment specialist at Educational Testing Service (ETS), Princeton, New Jersey, USA from 2000 to 2018, specializing primarily in the development of TOEFL iBT Listening tests.