Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test

Pablo Garcia Gomez *Educational Testing Service*
Aris Noah *Educational Testing Service*
Mary Schedl *Educational Testing Service*
Christine Wright *Educational Testing Service* and
Aline Yolkut *Educational Testing Service*

Providing information to test takers and test score users about the abilities of test takers at different score levels has been a persistent problem in educational and psychological measurement (Carroll, 1993). Since the 1990s Educational Testing Service has been investigating solutions to this problem through the development of proficiency scaling procedures and question-difficulty research. In 1997 a proficiency scale was developed for the Test of English as a Foreign Language (TOEFL) Reading Comprehension section using a tree-based regression approach. The current study describes a scale-anchoring study of the new TOEFL iBT reading test and the resulting proficiency descriptors that are now part of the TOEFL iBT score report. The goal was to provide descriptive information about the abilities that test takers need in order to answer questions correctly. These abilities are those articulated in the new TOEFL Reading Framework and in the guidelines for writing test questions. Scale anchoring is a method of creating descriptors of the performance of test takers that is based on both empirical data and judgments by test developers. It has been used with a variety of assessments, including the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS).

I Introduction

Providing information to test takers and test score users about the abilities of test takers at different score levels has been a persistent problem in educational and psychological measurement (Carroll, 1993). Since the 1990s Educational Testing Service has been investigating solutions to this problem through the development of proficiency scaling procedures (see, for example, Tatsuoka, Birenbaum, Lewis, and

Address for correspondence: Mary Schedl, Educational Testing Service, Rosedale Rd., MS 44-N, Princeton, NJ 08541, USA; email: mschedl@ets.org

Language Testing 2007 24 (3) 417–444 10.1177/0265532207077209

Sheehan, 1993 and Sheehan, 1997) and question-difficulty research (Kirsch and Mosenthal, 1990; Freedle and Kostin, 1993; Freedle, 1997; Nissan, De Vincenzi, and Tang, 1996). In 1997 a proficiency scale was developed for the Test of English as a Foreign Language (TOEFL) Reading Comprehension section (Sheehan, Ginther, and Schedl, 1999) using the tree-based regression approach (described in Sheehan, 1997). The current study describes a scale-anchoring study of the new TOEFL iBT reading test and the resulting proficiency descriptors that are now part of the TOEFL iBT score report.

II Background

Many lists of skills have been assembled over the years by both first and second language specialists based on the assumption that reading comprehension includes a number of different subskills or abilities (e.g. Bloom et al., 1956; Davis, 1968; Munby, 1978). Often these skills have been further classified as "higher-level" or "lower-level" skills (e.g. Barrett, 1968; Davies and Widdowson, 1974). However, empirical studies have failed to find evidence of separate subskills (Lunzer et al., 1979) or have provided contradictory evidence as to whether particular skills are distinct (Alderson, 2000).

Schedl et al. (1996) investigated the dimensionality of the reading comprehension subpart of the TOEFL reading section.[1] Performance on items classified as "reasoning items" was compared to performance on all other reading items in the test to determine whether reasoning items could be shown to measure a unique ability in addition to general reading ability. The authors concluded that differences in question types (such as understanding vocabulary, understanding factual information, making inferences vs. extrapolating information, making analogies, understanding organization and purpose, and understanding author's purpose/attitude) do not, in themselves, account for differences in difficulty or dimensionality.

In addition to the question of whether evidence can be found for separate subskills is the question of whether specialists can agree about which subskills are involved in particular reading tasks. A number of studies have failed to find agreement among specialists as to which subskills are being measured by reading items (Alderson, 1990a, 1990b; Alderson and Lukmani, 1989). Lumley (1993), how-

[1]At the time of the study the reading section consisted of two subparts, vocabulary and reading comprehension.

ever, found impressive agreement among English for Academic Purposes teachers in matching individual test items to specific subskills by following a procedure involving discussion and careful definition of terms. Recently a major effort was undertaken to develop a common European framework of references and a scale to describe language proficiency (North, 2000).

III Describing performance

A major goal of the TOEFL reading redesign (Enright, et al., 2000) was to provide performance descriptors to test takers that would help them interpret their test performance. To the extent that reading subskills required by particular reading tasks could be identified, interpretive information about the performance of individual test takers could be provided. Since research had shown that the abilities needed to answer test questions are not directly related to question type, reading tasks would have to be analyzed for the abilities necessary to perform them well, and differences in the difficulty of questions of the same type would have to be accounted for in terms of these abilities.

Using text and task variables to describe factors that cause difficulty for readers was seen to have potential as a means of providing feedback to test takers about their performance in a way that is more directly related to their abilities than a norm-referenced numerical score, which is only informative about a given test taker in comparison to others in the test-taking population.

Research by Irwin Kirsch and Peter Mosenthal (1990) indicated the importance of task variables in constructing a model of prose literacy that could account for performance differences on questions testing adult literacy. Their model of prose literacy as well as the text-comprehension models of Kintsch (1993, 1998), Gernsbacher (1990), and Mosenthal (1996) informed the analysis of the difficulty of 518 TOEFL paper-and-pencil reading questions (Sheehan, Ginther and Schedl, 1999). Difficulty variables derived from the latter study were used to code prototype questions for the first of two field studies carried out for TOEFL iBT in 2001. Difficulty variables for new types of tasks that did not exist in the paper-and-pencil test were hypothesized, and all questions were coded by test development staff. Training and discussion was extensive and care was taken to define and exemplify abilities relevant to the design of items, assuming this would result in greater agreement, as it had done for Lumley (1993).

An analysis of questions in a second field study (2003) was conducted by a group of test development and statistical analysis staff.

Two forms of the test were administered, and ten different types of test questions were included. Questions were sorted by difficulty for further analysis without regard to which passages they were associated with and without regard to the type of question. All of the easy and difficult questions were analyzed by TOEFL test developers using expert judgment to hypothesize the characteristics driving difficulty. Statements that describe the abilities thought to be required by easy tasks and by difficult tasks were drafted. These descriptions focused on the cognitive demands the questions made on the test taker across different types of test questions rather than on descriptions of the types of questions themselves.

A special pretesting event was conducted in 2004 for the purpose of acquiring equating questions for the new test, and another 24 sets of reading questions were coded. This kind of analysis is iterative and data driven. As such it is likely to go on for several years before providing sufficiently reliable information to serve as the basis of score reporting.

In order to provide more immediate test-taker feedback, a scale-anchoring study of 24 sets of questions that were pretested in 2003–2004 was carried out. A major difference between scale anchoring and difficulty coding as the source of interpretive information is that scale anchoring is carried out based on known difficulty and discrimination data, whereas difficulty coding is intended to predict and account for difficulty in advance. Another difference is that scale anchoring provides information about *typical* abilities of test takers at different points on the scale, whereas difficulty coding may ultimately make it possible to provide individual performance information. Test developers who participated in the scale-anchoring study were familiar with the difficulty variables associated with TOEFL iBT reading questions and used these as a basis for analyzing tasks that anchored to the scale. The scale-anchoring study both confirmed and extended our understanding of reading abilities. The descriptive performance information provided with the TOEFL iBT score report was derived from this scale-anchoring study (see Appendix A for the score report descriptors). A similar study was carried out for the TOEFL iBT listening test.

IV Scale anchoring: The process

As mentioned previously, one goal of TOEFL iBT was to make enhanced score reports available to test takers in order to provide them with more than just a number on a score scale. The goal was to provide descriptive information about the abilities that test takers need in order to

answer questions correctly. These abilities are those articulated in the Reading Framework (Enright et al., 2000) and in the guidelines for writing test questions. In creating test questions, test developers consider these abilities and select passages and develop sets of questions according to specifications delineating how many questions should be created to measure each of the abilities. In addition, test developers code variables related to difficulty for each question in the reading test.

In the design of the new enhanced score report, the initial goal was to create individualized descriptors using a statistical model based on MCMC (Monte Carlo Multiple Chains) and a software package called Arpeggio. However, this methodology is not yet mature enough to be used for score reports for TOEFL iBT, so it was decided to adopt a scale-anchoring approach instead. Scale anchoring is a method of creating descriptors of the performance of test takers that is based on both empirical data and judgments by test developers. It has been used with a variety of assessments, including the National Assessment of Educational Progress (NAEP) (Phillips et al., 1993; Jaeger, 2003) and the Trends in International Mathematics and Science Study (TIMSS).

Before examining the empirical data, a number of decisions had to be made:

- How many proficiency levels should there be?
- At what point on the score scale should each level begin and end?
- What criteria should be used to determine whether a test question "anchors" at a given level?

After consulting with colleagues who work on NAEP, staff in Assessment Development and Statistical Analysis decided that three levels would be adequate as a starting point and that the score scale would be divided into three equal percentiles. For a question to be considered an anchor at the High or Intermediate level, it had to meet three criteria:

- More than 50% of the people scoring at a given level (the conditional P value) had to answer the question correctly (a measure of difficulty).
- Fewer than 50% of the people scoring at a lower level answered the question correctly.
- The conditional P value at the next level down had to be at least 20 percentage points lower (a measure of discrimination).

For a question to be considered indicative of performance at the Low level, it had to meet only one formal criterion:

- At least 50% of the people at the Low level answered the question correctly.

Table 1 Some sample data

TOEFL IBT Reading Pretest data

QUESTION	N	OVERALL P+	LOW P+	INTERMEDIATE P+	D(20.00)	HIGH P+	D(20.00)
VB533037001	1460	54.52	24.39	56.26	*	87.39	
VB533038001	1474	66.89	38.93	74.13	*	91.95	
VB533039001	1476	77.03	51.93	85.68		97.46	
VB533041001	1474	78.97	52.21	89.78		99.15	
VB533424001	1470	53.47	36.16	47.92		78.77	*

Shaded cells indicate the level at which the question anchors.
* indicates that a question anchors at a given level.

In addition, staff considered informally the difference in the conditional P values of the Low and Intermediate levels. If the P values were similar for the two levels, the question was not considered in the analysis.

For example, in Table 1, the first question (VB533037001) anchors at the Intermediate level because 56.26% of the test takers at that score level answered correctly, while only 24.39% of the test takers at the lower level got it right. Similarly, the last question in Table 1 (VB533424001) anchors at the High level because 78.77% of the test takers at that score level answered it correctly, while only 47.92% of those scoring in the middle third of the score scale got it right. Because there are only three levels, no questions are said to anchor at the lowest level. However, in the discussion of the data for the Low level, both the third and the fourth questions in Table 1, and all similar items, were considered in the analysis, as more than 50% of the test takers in the Low level answered them correctly, and the conditional P values at the Intermediate level are considerably higher than those at the Low level.

Once these decisions were made and the analyses run, the judgment part of the anchoring analysis began.[1] The statistical data were merged with the text of the test questions and the passage, and this information was divided into four categories: those questions that anchor at the High level, those that anchor at the Intermediate level, those considered indicative of skills at the Low level, and the rest of the questions, which did not anchor at any level. A group of Reading test developers internal to ETS (the authors of this paper) was convened, and this group worked to articulate the knowledge, skills, and abilities that were demonstrated by correct responses to the questions at each level. The

[1]The data for the scale anchoring came from 24 Reading sets pretested in 2003–2004. Data were available for ~312 test questions.

next section of this paper provides exemplar questions and the associated descriptive text. Although there were no data relevant to the reading passage itself, care was taken to consider the characteristics of the text as well as the abilities needed to answer the question correctly.

After completing the analysis of the questions at each level, the group turned its attention to the creation of concise descriptors of overall performance at each level. These descriptors can be found in Appendix A. It should be noted that these descriptors encapsulate what test takers at a given level typically are able to do; they are not meant to be descriptive of an individual's performance.

V Analysis of test questions from the scale-anchoring study

Test questions in TOEFL iBT reading focus on the measurement of abilities needed to read for two major academic purposes. The questions measuring *basic comprehension* primarily assess lexical, syntactic, and semantic abilities along with the abilities to understand information presented in single sentences and to connect information across sentences. The questions measuring *reading to learn* require more than understanding discrete points and getting the general idea based on the lexical, syntactic, and semantic content of texts. Reading-to-learn test questions assess specific abilities that contribute to learning: recognizing the organization and purpose of the text, distinguishing major from minor ideas and essential from nonessential information, understanding rhetorical functions (such as cause-effect relationships, compare-contrast relationships, and arguments), and conceptualizing and organizing text information into a mental framework. Having an organized mental representation of the text is seen as critical to learning from the text because it allows the reader to remember important information and apply it in new situations (Enright et al., 2000; Enright and Schedl, 2000).

In this section, a sampling of basic comprehension and reading-to-learn questions from the study that anchored at each of the three levels of performance are presented and analyzed in relation to the ability descriptors and the factors that determine their degree of difficulty. Test questions measuring basic comprehension typically take the form of a "stem" that poses the query of the test question and indicates the part of the text, the targeted text (typically a paragraph), that is relevant to answering the query. The stem is followed by four options from which the test taker must select the correct answer to the query. The format of questions measuring reading-to-learn abilities

varies according to whether a table or a summary of the passage is used. It should be added that test takers respond to all questions only after reading the entire passage, not in isolation from the broader context as the questions appear in this paper. The reader of this paper is therefore encouraged to read the full passages (available in Appendix B) and become familiar with the broader context of the questions included in the discussion that follows.

VI Factual information test questions

TOEFL iBT Factual Information questions play a central role in measuring the ability of test takers to read a text for basic comprehension. The sets of questions used in the scale-anchoring study included a large number of Factual Information questions, and analysis of the skills required to give correct responses to these and other basic comprehension questions led to conclusions both about test takers' abilities at different levels of performance and about the factors that make some of the questions more difficult than others. These conclusions are summarized in the ability descriptors for the Low, Intermediate, and High levels of performance (Appendix A).

Based on the analyses, test takers at the Low level (the lowest third of the test-taking population) have limited ability to understand individual sentences and to connect information across two or more sentences. Test takers with this limited level of comprehension typically have difficulty answering a Factual Information question correctly when the correct answer—and/or the stem—of the question involves significant paraphrasing and hence cannot be easily matched to the text. This is because Low-level test takers rely heavily on particular words and phrases in order to identify the correct answer to a question. This strategy can be successful when the area of the text targeted for testing is relatively straightforward, but the strategy does not work well when the text combines difficult vocabulary and syntax with complexity of concepts.

At the Intermediate level, test takers have greater ability to understand individual sentences and can connect information across two or more sentences. They also have considerable ability to recognize information even when it is significantly paraphrased, and hence they rely less on matching words and phrases to the text. Intermediate-level test takers can use these abilities to give correct responses to Factual Information questions as long as the targeted text does not use very low-frequency vocabulary and its conceptual complexity is not too great.

What distinguishes test takers at the High level is that their ability to understand and connect information and to recognize paraphrased information allows them to answer Factual Information questions correctly even when the targeted text has very low-frequency vocabulary and has a high level of conceptual density.

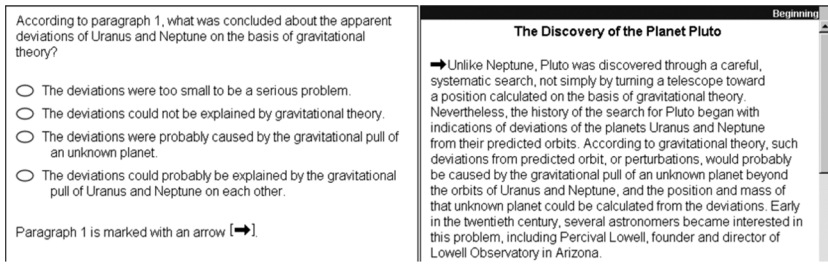
The term "conceptual density" requires some explanation. In general, when comparing two specific texts, it is not difficult to judge whether one text is more or less conceptually dense than another: for instance, a portion of text that consists of a general statement followed by a series of examples is typically less conceptually dense—less densely packed with meaning—than one that develops a systematic comparison and contrast between two views, theories, or artistic styles. But because a great variety of elements influence a text's conceptual density, it is difficult to provide a definition. In general, a text can be said to have a relatively high degree of conceptual density when it requires the reader to carry over and assemble—often not all at once but in stages—the meaning of concepts that build upon each other through complex interrelationships. The greater the number of conceptual interrelationships and the more complex they are, the more conceptually dense a text is likely to be. The types of conceptual interrelationships may vary greatly from one text (or part of a text) to another. Factual Information, Rhetorical Purpose, and Reading-to-learn questions presented in the discussion that follows include examples that test conceptually dense text.

Four examples will help illustrate test-taker performance on Factual Information questions at the three levels.

Performance at the low level Test takers at the Low level have the ability to understand information that is of limited complexity in texts that are relatively straightforward, and they can respond correctly to questions targeting this type of text when the stem and correct answer resemble the text closely or are simple paraphrases of it. In Question 1 from the passage *The Discovery of the Planet Pluto*, 57.68% of the lowest-third of the test-taking population answered the question correctly.

Relevant text: Nevertheless, the history of the search for Pluto began with indications of deviations of the planets Uranus and Neptune from their predicted orbits. According to gravitational theory, such deviations from predicted orbit, or perturbations, would probably be caused by the gravitational pull of an unknown planet beyond the orbits of Uranus and Neptune, and the position and mass of that unknown planet could be calculated from the deviations.

Question 1



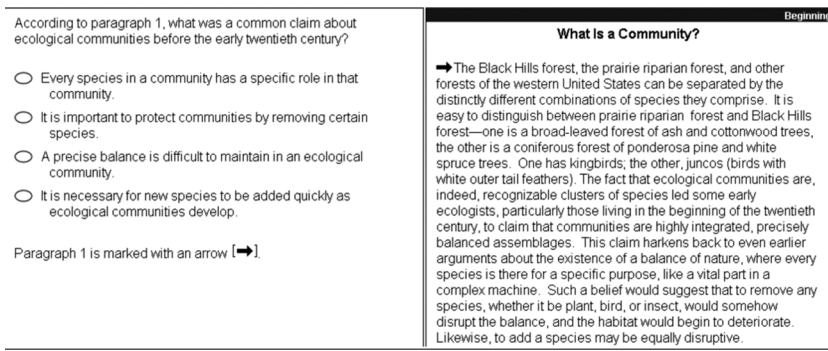
Stem: According to paragraph 1, what was concluded about the apparent deviations of Uranus and Neptune on the basis of gravitational theory?

Correct answer: The deviations were probably caused by the gravitational pull of an unknown planet.

Note the underlined areas of overlap between the text, stem, and correct answer. The majority of test takers at the Low ability level were able to perform this task. In this example, note that the stem and correct answer use the exact words and phrases of the text and that the task did not require a significant degree of connecting information across sentences. For Intermediate-level test takers, this task presented less of a challenge: 87.16% of them answered correctly.

Performance at the intermediate level Test takers at the Intermediate level of performance are able to connect information across two or more sentences and can recognize information from the text even when that information does not match the wording of the text. Test takers at this level demonstrate the ability to answer correctly even when the tar-

Question 2



geted text area contains relatively difficult vocabulary and a complex grammatical structure. Two examples of questions that anchored at this level from the passage *What Is a Community?* will help illustrate the abilities of Intermediate-level test takers. In the first example (Question 2), 65.5% of test takers at this level answered correctly, while only 28% of the Low-level test takers answered correctly.

Relevant text: The fact that ecological communities are, indeed, recognizable clusters of species led some early ecologists, particularly those living in the beginning of the twentieth century, to claim that communities are highly integrated, precisely balanced assemblages. This claim harkens back to even earlier arguments about the existence of a balance of nature, where every species is there for a specific purpose, like a vital part of a complex machine.

Stem: According to paragraph 1, what was a common claim about ecological communities before the early twentieth century?

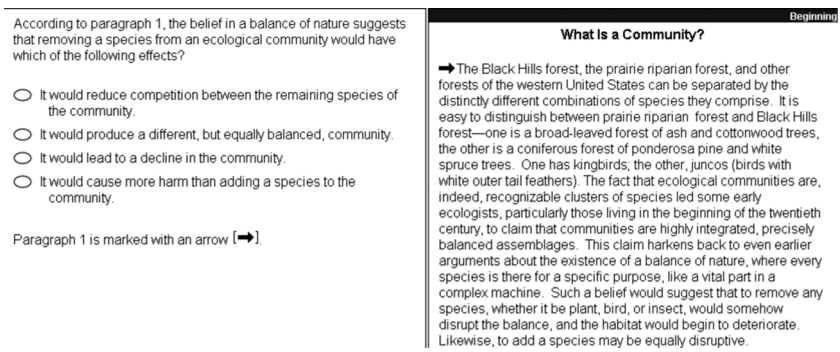
Correct answer: Every species in a community has a specific role in that community.

A correct response to this question requires more than matching phrases. The test taker must connect what is said about the twentieth century to "even earlier arguments" and then identify the correct answer by recognizing information about the purpose of every species in a relatively simple paraphrase ("role" is used in the answer instead of "purpose").

A second example (Question 3) from the same paragraph of the passage illustrates another aspect of the abilities exhibited by test takers performing at the Intermediate level

Relevant text: This claim harkens back to even earlier arguments about the existence of a balance of nature, where every species is there for a

Question 3


According to paragraph 1, the belief in a balance of nature suggests that removing a species from an ecological community would have which of the following effects?

- It would reduce competition between the remaining species of the community.
- It would produce a different, but equally balanced, community.
- It would lead to a decline in the community.
- It would cause more harm than adding a species to the community.

Paragraph 1 is marked with an arrow [➡].

Beginning

What Is a Community?

➡The Black Hills forest, the prairie riparian forest, and other forests of the western United States can be separated by the distinctly different combinations of species they comprise. It is easy to distinguish between prairie riparian forest and Black Hills forest—one is a broad-leaved forest of ash and cottonwood trees, the other is a coniferous forest of ponderosa pine and white spruce trees. One has kingbirds; the other, juncos (birds with white outer tail feathers). The fact that ecological communities are, indeed, recognizable clusters of species led some early ecologists, particularly those living in the beginning of the twentieth century, to claim that communities are highly integrated, precisely balanced assemblages. This claim harkens back to even earlier arguments about the existence of a balance of nature, where every species is there for a specific purpose, like a vital part in a complex machine. Such a belief would suggest that to remove any species, whether it be plant, bird, or insect, would somehow disrupt the balance, and the habitat would begin to deteriorate. Likewise, to add a species may be equally disruptive.

specific purpose, like a vital part in a complex machine. Such a belief would suggest that to remove any species, whether it be plant, bird, or insect, would somehow disrupt the balance, and the habitat would begin to deteriorate. Likewise, to add a species may be equally disruptive.

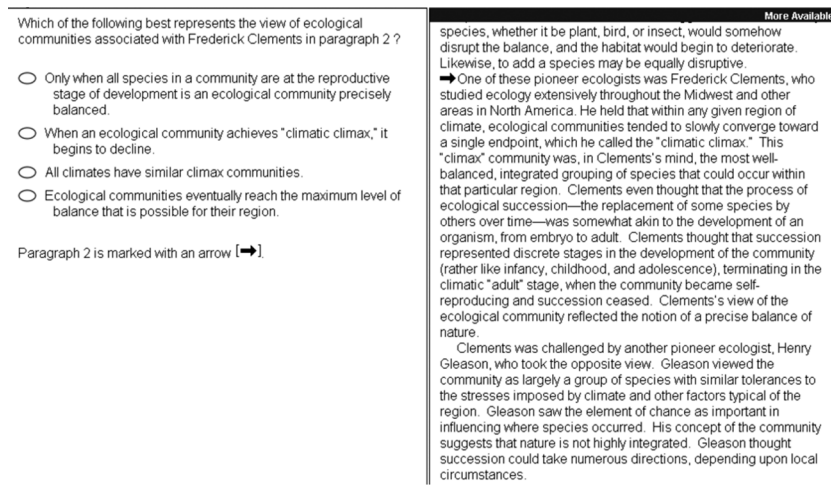
Stem: According to paragraph 1, the belief in a balance of nature suggests that removing a species from an ecological community would have which of the following effects?

Correct Answer: It would lead to a decline in the community.

Over 70% of Intermediate-level test takers were able to answer this question correctly compared to 25% of Low-level test takers. This question does not require a significant ability to connect information: test takers need only connect "a balance of nature" to the phrase "such a belief" to understand why, according to this belief, removing a species would have a negative effect. This question, however, requires a significant ability to recognize paraphrased information and depends on the understanding of vocabulary: "a decline in the community" in the answer is not easy to match to the wording in the text ("the habitat would begin to deteriorate").

Performance at the high level A Factual Information question (Question 4) targeting the next paragraph of *What Is a Community?* illustrates the abilities of test takers at the High level of performance: 74.4% of High-level test takers answered this question correctly compared to only 36.13% of Intermediate-level test takers.

Question 4



Relevant Text: One of these pioneer ecologists was Frederick Clements, who studied ecology extensively throughout the Midwest and other areas in North America. He held that within any given region of climate, ecological communities tended to slowly converge toward a single endpoint, which he called the "climatic climax." This "climax" community was, in Clements's mind, the most well-balanced, integrated grouping of species that could occur within that particular region.

Stem: Which of the following best represents the view of ecological communities associated with Frederick Clements in paragraph 2?

Correct answer: Ecological communities eventually reach the maximum level of balance that is possible for their region.

The targeted text, the whole of paragraph 2, in Question 4 formulates a number of different ideas that Clements held about ecological communities, but the stem does not help the examinee locate which of these ideas is the one expressed in the correct answer. As a result, the test taker must understand all of Clements' ideas about ecological communities in order to reject incorrect answers and identify the correct answer. The task of processing the ideas of this entire paragraph, however, presents a considerable challenge to test takers because of the target text's high degree of conceptual density.

The two examples of Factual Information questions from *What Is a Community?* that anchored at the Intermediate level targeted paragraph 1, whereas Question 4 targets paragraph 2. Paragraph 1 serves as a general introduction to the idea of "a balance of nature." In paragraph 2, however, the concept of "the most well-balanced community" is expanded and refined. Test takers must understand three more dimensions of that concept as it was elaborated by Clements: (1) that "the most well-balanced community" differs from region to region, (2) that an ecological community in a given region evolves through time toward its own ultimate state of balance, and (3) that this developmental process is analogous to an organism's development from infancy to adulthood. Note that each of the three incorrect answers is designed to test the comprehension of one of these three added dimensions.

In addition, the correct answer represents a significant paraphrase of the wording in the text ("eventually reach" for "slowly converge" and "the maximum level of balance possible" for "the most well-balanced, integrated grouping of species that could occur").

The interplay of conceptually dense text, a stem that does not help the test taker locate the necessary information to answer correctly, and an answer that is significantly paraphrased resulted in a question that proved difficult for all but the most able of TOEFL test takers.

VII Rhetorical purpose test questions

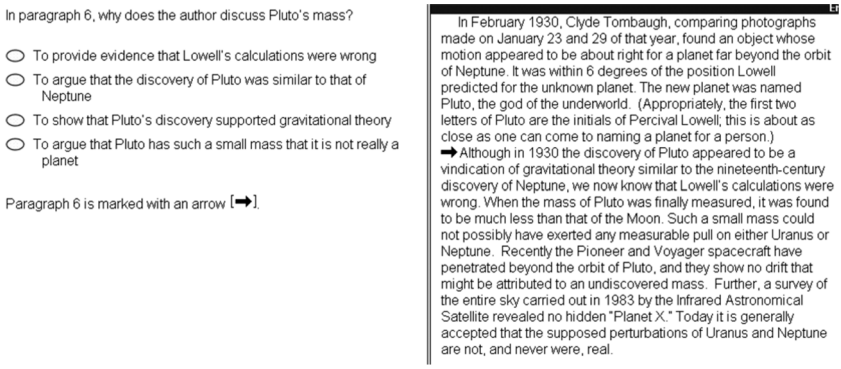
TOEFL iBT Rhetorical Purpose questions test the ability to recognize the expository organization of a portion of text and the role specific information serves within that text.

Performance at the Low level The descriptors that emerged from our analyses state that test takers at the Low performance level "have difficulty identifying the author's purpose except when that purpose is explicitly stated in the text or easy to infer from the text." This statement implies that Low-level test takers *are able* to identify the author's purpose when it is explicitly stated or easy to infer from the text, and indeed there are Rhetorical Purpose questions in the scale-anchoring study that have been answered correctly by more than 50% of Low-level test takers. Test-security concerns prevent the publication at this time of specific examples for this Low ability level, so a general description of such test questions follows.

A question of this type asks why some word or phrase is mentioned in a certain portion of text; e.g. "In paragraph 3, why does the author mention dinosaurs?" The relevant text is centered on the computer screen with the word "dinosaurs" highlighted within the text. The paragraph in which the term "dinosaurs" is mentioned has an expository structure quite common in TOEFL iBT reading passages (and introductory college textbooks in general). It begins with a general statement; e.g., "Many species became extinct as a result of catastrophic impacts by asteroids or meteorites," directly followed by a relatively straightforward list of examples, with the extinction of the dinosaurs serving as one of the examples. The text helps the reader understand that the author's purpose is to illustrate the general statement by examples. This help may come in two forms: either the series of examples is marked by an *explicit indicator* of rhetorical purpose (such as "for example" or "for instance") or the series is located directly after the general statement, so that the reader can easily infer (from the *location*) that the author's purpose in mentioning the extinction of the dinosaurs is to illustrate the general statement. Finally, the correct answer is worded simply and straightforwardly ("To provide an example of species that became extinct as a result of asteroid or meteorite impacts").

When the author's purpose is not explicitly indicated in the text or easy to infer from the text (from location and/or other cues), and especially when the series of examples is not a simple list but is interspersed with comments, qualifications, etc., test takers at the Low level of performance do not as a group answer correctly.

Question 5



Performance at the Intermediate level The descriptors that emerged from our analyses state that test takers at the Intermediate performance level "can recognize the expository organization of a text and the role that specific information serves within a larger text but have some difficulty when these are not explicit or easy to infer from the text," while test takers at the High level can recognize text organization and the role served by specific information "even when the text is conceptually dense." The claim is that conceptually dense texts make it difficult for Intermediate level test takers to infer rhetorical purpose that is not explicitly marked.

Question 5 is an example of a Rhetorical Purpose question (on *The Discovery of the Planet Pluto*) that anchored at the Intermediate level (The correct answer is the first option: 67.27% of test takers at the Intermediate level answered correctly, but only 37.81% of those at the Low level answered correctly). This test question is based on a potentially more complex type of expository structure than the one described previously: instead of a general statement followed by illustrations, what we have here is evidence presented in support of (or against) a view, hypothesis, thesis, claim, etc.

Note that the author's purpose in discussing Pluto's mass is not marked in the text by an explicit indicator and thus needs to be inferred. Moreover, the paragraph tested has a degree of conceptual density that makes it difficult for test takers at the Low level (but not for those at the Intermediate level) to infer the author's purpose from the location of the term "Pluto's mass" and/or other cues.

The paragraph tested is conceptually denser than the type of paragraph described in the dinosaur example. The thesis or claim for which evidence is to be offered is that "we now know that Lowell's

calculations were wrong" (even though, as the previous paragraph states, Pluto was discovered close to a position previously predicted by those calculations). That thesis, however, comes after an "although" clause that introduces a conceptual distinction between what appeared and what was, in fact, the case. (Options 2 and 3 are attractive if that distinction is not noted.) After the thesis is stated, the topic of Pluto's mass is introduced in the next sentence, but the relevance of the topic to the thesis becomes clear in subsequent sentences (hence, the location of the term "Pluto's mass" does not help narrow the search for the answer). An additional complication is the discussion of "Planet X," the relevance of which to the discussion of Pluto's mass is not a matter of explicit statement but of reader's inference.

Performance at the High level An even higher degree of conceptual density can be illustrated by Question 6, a Rhetorical Purpose test question (on *What Is a Community?*) of the same general kind as the previous one but more complex, which anchored at the High level (The correct answer is the first option: 71.27% of test takers at the High level answered correctly, but only 45.14% of those at the Intermediate level and 27.06% of those at the Low level answered correctly). Faced with this degree of conceptual density, most test takers at the Intermediate level were unable to infer the author's purpose correctly, whereas most of those at the High level were able to do so.

Note the conceptual sophistication of the text and, in particular, the complexity of the formulations of the Clements and Gleason views. This is the main factor that accounts for the high difficulty of this Rhetorical Purpose test (Question 6). Instead of having supporting evidence for a single thesis as in the previous example, here we have three different views being considered: Clements', Gleason's, and "the current view."

Question 6

In paragraph 5, why does the author mention green ash trees and plains cottonwood trees?

- ○ To support the current view of the relations among members of an ecological community
- ○ To provide examples of species that prefer to live on floodplains
- ○ To provide evidence that supports the theory of Clements
- ○ To show where one ecological community stops and the other begins

Paragraph 5 is marked with an arrow [➡].

Who was right? Many ecologists have made precise measurements, designed to test the assumptions of both the Clements and Gleason models. For instance, along mountain slopes, does one life zone, or habitat type, grade sharply or gradually into another? If the divisions are sharp, perhaps the reason is that the community is so well integrated, so holistic, so like Clements viewed it, that whole clusters of species must remain together. If the divisions are gradual, perhaps, as Gleason suggested, each species is responding individually to its environment, and clusters of species are not so integrated that they must always occur together.

➡It now appears that Gleason was far closer to the truth than Clements. The ecological community is largely an accidental assemblage of species with similar responses to a particular climate. Green ash trees are found in association with plains cottonwood trees because both can survive well on floodplains and the competition between them is not so strong that only one can persevere. One ecological community often flows into another so gradually that it is next to impossible to say where one leaves off and the other begins. Communities are individualistic.

To recognize the first option as the correct answer, the reader needs to understand from the passage that the author is moving beyond an account of the two historical models to express the current view (which is closer to Gleason's but presumably not identical to it). The second sentence of the paragraph formulates the current view, and the third sentence supports that general statement with a factual statement about particular species.

Another notable feature is the second option of Question 6: it is true that green ash and plains cottonwood trees are examples of species "that prefer to live on floodplains," but it is not true that these trees are introduced in the paragraph in order to provide an example of such species. Having to draw a distinction between the truth of the second option's factual content and its failure to capture the author's rhetorical purpose is not easy for TOEFL test takers and may lend Rhetorical Purpose questions an extra dimension of difficulty.

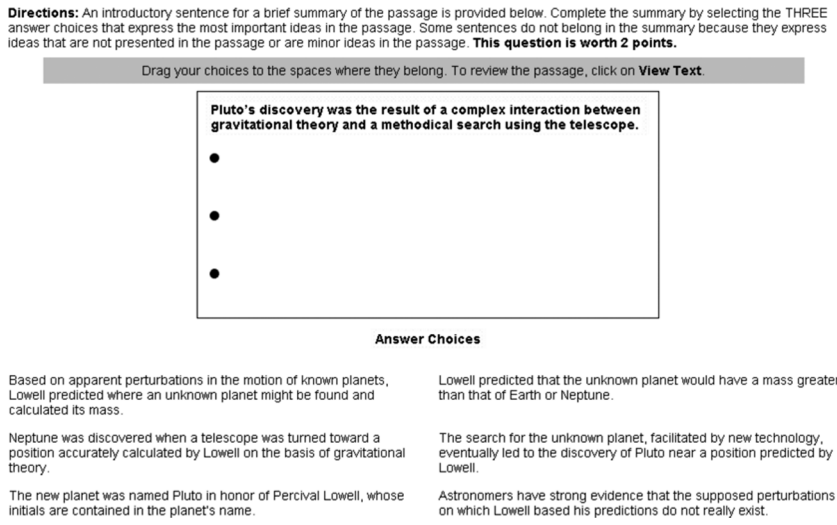
VIII Reading-to-learn test questions

The last section of the descriptors refers to the test takers' ability to abstract the major ideas from an academic text. Before reading this section of the paper, we encourage the reader to go over the passages *The Discovery of Planet Pluto* and *What is a Community?* and their respective summary test questions. Full-length copies of the passages can be found in Appendix B. Their corresponding summary questions are included in this section.

TOEFL iBT's reading measure includes new Reading-to-learn test question types that require test takers to understand the rhetorical pattern of a text as well as to connect and integrate the information contained in the passage into a coherent whole.

These new questions can take the form of a summary or a schematic table, and each reading set has one. The summary question consists of an introductory sentence and six options, three of which are the correct answers that, together with the introductory sentence, summarize the major ideas presented in the passage. Summary questions are worth two points and test takers can receive partial credit. To obtain one point, test takers must correctly identify two out of the three correct answers. To obtain full credit for this test question (two points), test takers must recognize all three correct answers. We studied the test takers' performance on these questions in order to evaluate their ability to abstract the major ideas from an academic text and in order to determine how that ability varied from level to level.

Question 7



Correct answers: 1, 4, 6 (Note: Option layout is 1 . . . 2
3 . . . 4
5 . . . 6)

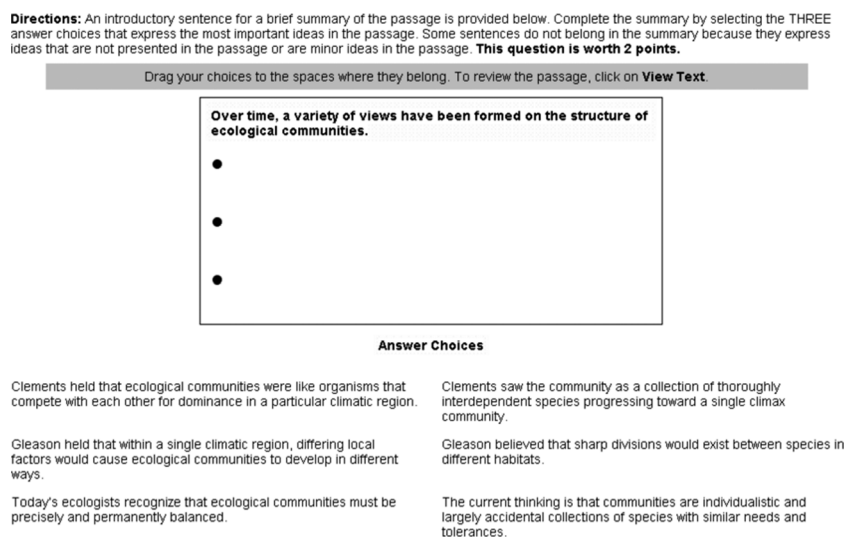
The scale-anchoring study showed that Low-level test takers can sometimes recognize major ideas from a text. However, they are only able to do this when the information is clearly presented, memorable, or illustrated by examples. Denser, more abstract texts pose difficulties for these test takers. More complex syntax and vocabulary, and/or text organization that is not strongly marked are also sources of difficulty at this level. The summary question from *The Discovery of Planet Pluto* illustrates what Low-level test takers can do: 62.15% of the test takers at the lower end of the scale were able to identify two of the three correct answers.

Although some parts of this passage are denser than others, the relationships among meaning units that shape most of the major ideas of the text are not complex and are explicitly stated in the passage. For example, for the first major point (option 1) there is a simple linear path: ". . . such deviations from predicted orbit, or perturbations [of Uranus and Neptune] would be caused by the pull of an unknown planet"; "the position and mass of that unknown planet could be calculated from the deviations"; Lowell did this. The second correct answer (option 4) involves a simple synthesis of the clearly illustrated

information presented in paragraphs 3, 4, and 5. The new "photographic telescope" and "the invention of the blink microscope" are combined in the correct answer as "new technology" that was used in the discovery of the planet in 1930. This whole process is also facilitated by the fact that the information conveyed is mainly factual and concrete and by the fact that the passage presents elements that are characteristic of the narrative genre. The third correct answer (option 6), however, involves processing a denser path, since it requires an understanding of how the information presented in the last paragraph challenges what was previously stated in the passage. Not surprisingly, this last correct answer proved to be the most difficult of the three. Nevertheless, it is important to note here that within the range of passages used in TOEFL iBT Reading, *The Discovery of Planet Pluto* is one of medium difficulty, and test takers at the lower end of the scale performed better when dealing with relatively easier passages.

Test takers at the Intermediate level are also able to abstract the major ideas from the text, and the task is clearly much easier for them when the text has the characteristics that allow those at the Low level to perform relatively well. For example, in the same summary question, 85.01% of the population at the Intermediate level correctly identified two out of the three correct answers. However, even though more complex syntax and vocabulary are less of a problem at the Intermediate level than at the Low level, test takers at the Intermediate level still have difficulty abstracting major ideas when the text is denser, that is, when the conceptual connections between meaning units in the building up of information is more complex. The passage *What is a Community?* and its summary question (Question 8) provide a good example. *What is a Community?* is a denser passage, mostly conceptual and abstract with overall harder syntax and vocabulary. Unlike *The Discovery of Planet Pluto*, in which certain parts are denser than others, in *What is a Community?* the conceptual density is relatively consistent throughout the text. The connections between meaning units required to successfully abstract the major points of this passage are greater in number and significantly more complex. For example, all three correct answers (options 2, 3, and 6) require test takers to clearly recognize the synthesis they present of the explanations of Clements' and Gleason's views, which cannot be fully understood separately but only in comparison with one another—in other words, Clements' views cannot be fully understood until Gleason's views are processed and vice versa. Of the Intermediate-level test takers, 64% recognized two of the correct answers—a significant difference compared with the previous example (85.01%)—and only 20% of

Question 8



Correct answers: 2, 3, 6

these test takers were able to receive full credit for this question by correctly identifying all three correct answers.

By contrast, test takers at the higher end of the scale are able to abstract the major ideas even when the text is dense and contains complex language. In the summary of *What is a Community?* (Question 8), 90.05% of the test takers at the High level correctly identified two of the correct answers, and 55% received full credit for this question.

The analysis of the test takers' ability to abstract major ideas from a text at the Low, Intermediate, and High levels showed the same pattern across all passages and reading-to-learn questions included in the analysis.

IX Outstanding issues and concerns

The potential for misuse and misunderstanding of the descriptors exists. Linn and Dunbar (1992) have described the confusion of the general public about the meaning of NAEP data related to score anchors. They conclude that the reasons for the discrepancy between the percentage of examinees who answer an anchor item correctly and the percentage who score above the corresponding anchor point may

be too subtle for mass communication. Phillips et al., (1993) describe the potential danger of overinterpreting examinee performance at anchor points so that all examinees at a particular level are assumed to be proficient at all abilities measured at that level. Even though the TOEFL descriptors of performance will not be sent to institutions without the permission of the individual test taker, it is clearly important to ensure that inferences based on these descriptors are sound.

The analysis presented in this paper leaves a number of additional issues and concerns to be addressed:

- Would the data support more than three levels? Descriptors for any given level work best for the typical test taker at that level. The further a test taker's score deviates from the middle of the range, the less well the descriptors for that range fit the individual test taker.
- Anchor questions for each level were selected across a large score range. Would it be more appropriate to select anchors from the middle of each range?
- Are the criteria for difficulty and discrimination rigorous enough? Too rigorous?
- Is there a better way to identify items at the Low level that are truly indicative of performance at that level?
- Although the pretest population was carefully selected to represent the TOEFL population in terms of language background, country of origin, gender, age, etc., it is, nevertheless, a pretest population. As such, this population may differ from the operational population in terms of motivation or familiarity with the types of questions typically included in TOEFL iBT. Would a second study, using data from an operational test, yield significantly different results?
- Would test developers and subject matter experts external to ETS characterize the abilities in ways similar to test developers internal to ETS?
- Are the descriptors useful to test takers, particularly at the Low and Intermediate levels?
- Although TOEFL iBT is designed to ensure comparability across different test forms, do these descriptors also truly generalize?
- Are there tools available for analyzing the text itself that might inform the process of characterizing test takers' abilities?

In the future, ETS plans to address each of these questions. A second scale-anchoring study that will address many of these questions began in late 2006, using worldwide operational test data. Statisticians and test developers will redefine examinee proficiency levels, replicating the

procedures and identifying items that distinguish among the levels in order to validate the descriptors of abilities at each level. Of special interest is the question of whether the particular cut scores and descriptors used in this study will discriminate operational test takers as successfully as they discriminated the test taker sample used in this study. If necessary, cut scores and descriptors will be revised to aid in the interpretation of abilities. It would also be appropriate to validate the descriptors using teachers or other test developers.

X References

Alderson, J.C. 1990a: Testing reading comprehension skills (Part one). *Reading in a Foreign Language* 6, 425–38.

——1990b: Testing reading comprehension skills (Part two). *Reading in a Foreign Language* 7, 465–503.

—— 2000: *Assessing reading*. New York: Cambridge University Press.

Alderson, J.C. and **Lukmani, Y.** 1989: Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language* 5, 353–70.

Barrett, T.C. 1968: What is reading? In Clymer, T., editors, *Innovation and change in reading instruction.* 67th Yearbook of the National Society for the Study of Education, University of Chicago Press.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. and **Kratwohl, D.R.,** editors, 1956: *Taxonomy of educational objectives: Cognitive domain.* New York: David McKay.

Carroll, J.B. 1993: Test theory and the behavioral scaling of test performance. In Fredericksen, N., Mislevy, R.J. and Bejar, I., editors, *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum, 297–322.

Davies, A. and **Widdowson, H.** 1974: The teaching of reading and writing. In Allen, J.P.B. and Corder, S.P., editors, *Techniques in applied linguistics*, Vol. 3. Oxford: Oxford University Press.

Davis, F.B. 1968: Research in comprehension in reading. *Reading Research Quarterly* 3, 499–545.

Enright, M.K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P. and **Schedl, M.** 2000: *TOEFL 2000 Reading Framework: A Working Paper.* TOEFL Monograph Series Report 17. Educational Testing Service.

Enright, M.K. and **Schedl, M.** 2000: *Reading for a reason: Using reader purpose to guide test design*. TOEFL Internal Report, Educational Testing Service.

Freedle, R. and **Kostin, I.** 1993: *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items.* TOEFL Research Report 44, Educational Testing Service.

Freedle, R. 1997: The relevance of multiple-choice reading test data in studying expository passage comprehension: The saga of a 15 year effort towards an experimental/correlational merger. *Discourse Processes* 23, 399–440.

Gernsbacher, M.A. 1990: *Language comprehension as structure building.* Hillsdale, NJ: Lawrence Erlbaum.

——1996: The structure-building framework: What it is, what it might also be, and why. In Britton, B.K. and Graesser, A.C., editors, *Models of text understanding*. Hillsdale, NJ: Erlbaum, 289–311.

—— 1997: Two decades of structure building. *Discourse Processes* 23, 265–304.

Jaeger, R.M. 2003: NAEP validity studies: *Reporting the results of the National Assessment of Educational Progress,* NCES. Washington, DC: National Center for Education Statistics, U.S. Department of Education, 11.

Kintsch, W. 1993: Information accretion and reduction in text processing: Inferences. *Discourse Processes* 16, 193–202.

——1998: *Comprehension: A paradigm for cognition.* Oxford: Cambridge.

Kirsch, I.S. and **Mosenthal, P.B.** 1990: Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly* 25, 5–30.

Linn, R.L. and **Dunbar, S.** 1992: Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement* 29, 177–94.

Lumley, T. 1993: The notion of subskills in reading comprehension tests: An EAP example. *Language Testing* 10, 211–34.

Lunzer, E., Waite, M. and **Dolan, T.** 1979: Comprehension and comprehension skills. In Lunzer, E. and Garnder, K., editors, *The effective use of reading.* London: Heinemann Educational, 37–71.

Mosenthal, P.B. 1996: Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology* 88, 314–32.

Munby, J. 1978: *Communicative syllabus design.* Cambridge: Cambridge University Press.

Nissan, S., De Vincenzi, F. and **Tang, K.L.** 1996: *An analysis of factors affecting the difficulty of dialogue items in TOEFL listeningcomprehension*. TOEFL Research Report 51, Educational Testing Service.

North, B. 2000: *The development of a common framework scale of language proficiency.* New York: Peter Lang.

Phillips, G.W., Mullis, I.V.S., Bourque, M.L., Williams, P.L., Hambleton, R.K., Owen, E.H. and **Barton, P.E.** 1993: *Interpreting NAEP scales*, NCES 93421. Washington, DC: National Center for Education Statistics, US Department of Education.

Schedl, M., Gordon, A., Carey, P.A. and **Tang, L.T.** 1996: *An analysis of the dimensionality of TOEFL reading comprehension items.* TOEFL Research Report 53, Educational Testing Service.

Sheehan, K.M. 1997: A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement* 34, 333–52.

Sheehan, K.M., Ginther, A. and **Schedl, M.** 1999: *The development of a proficiency scale for the TOEFL reading comprehension section*. Paper presented at the Annual Conference of the Association for Applied Linguistics, Stamford, CT (March, 1999).

Tatsuoka, K., Birenbaum, M., Lewis, C. and **Sheehan, K.** 1993: *Proficiency scaling based on conditional probability functions for attributes*. ETS Research Report RR-93–50-ONR, Educational Testing Service.

Appendix A: TOEFL iBT score report descriptors

Reading Skills	Level	Your Performance
Reading	High (22–30)	Test takers who receive a score at the **HIGH** level, as you did, typically understand academic texts in English that require a wide range of reading abilities regardless of the difficulty of the texts. Test takers who score at the **HIGH** level typically • have a very good command of academic vocabulary and grammatical structure; • can understand and connect information, make appropriate inferences, and synthesize ideas, even when the text is conceptually dense and the language is complex; • can recognize the expository organization of a text and the role that specific information serves within the larger text, even when the text is conceptually dense; and • can abstract major ideas from a text, even when the text is conceptually dense and contains complex language.
Reading	Intermediate (15–21)	Test takers who receive a score at the **INTERMEDIATE** level, as you did, typically understand academic texts in English that require a wide range of reading abilities, although their understanding of certain parts of the texts is limited. Test takers who receive a score at the **INTERMEDIATE** level typically • have a good command of common academic vocabulary but still have some difficulty with high-level vocabulary; • have a very good understanding of grammatical structure; • can understand and connect information, make appropriate inferences, and synthesize information in a range of texts but have more difficulty when the vocabulary is high level and the text is conceptually dense; • can recognize the expository organization of a text and the role that specific information serves within a larger text but have some difficulty when these are not explicit or easy to infer from the text; and • can abstract major ideas from a text but have more difficulty doing so when the text is conceptually dense.

Reading	Low (0–14)	Test takers who receive a score at the **LOW** level, as you did, typically understand some of the information presented in academic texts in English that require a wide range of reading abilities, but their understanding is limited. Test takers who receive a score at the **LOW** level typically • have a command of basic academic vocabulary, but their understanding of less common vocabulary is inconsistent; • have limited ability to understand and connect information, have difficulty recognizing paraphrases of text information, and often rely on particular words and phrases rather than a complete understanding of the text; • have difficulty identifying the author's purpose, except when that purpose is explicitly stated in the text or easy to infer from the text; and • can sometimes recognize major ideas from a text when the information is clearly presented, memorable, or illustrated by examples but have difficulty doing so when the text is more demanding.

Appendix B*

What is a community?

The Black Hills forest, the prairie riparian forest, and other forests of the western United States can be separated by the distinctly different combinations of species they comprise. It is easy to distinguish between prairie riparian forest and Black Hills forest—one is a broad-leaved forest of ash and cottonwood trees, the other is a coniferous forest of ponderosa pine and white spruce trees. One has kingbirds; the other, juncos (birds with white outer tail feathers). The fact that ecological communities are, indeed, recognizable clusters of species led some early ecologists, particularly those living in the beginning of the twentieth century, to claim that communities are highly integrated, precisely balanced assemblages. This claim harkens

*TOEFL iBT test specifications require that passages such as *What is a Community?* and *The Discovery of the Planet Pluto* be excerpted from published works such as college-level textbooks or books of general academic interest. Passages may occasionally contain minor revisions to make the excerpted material self-contained and suitable for testing.

back to even earlier arguments about the existence of a balance of nature, where every species is there for a specific purpose, like a vital part in a complex machine. Such a belief would suggest that to remove any species, whether it be plant, bird, or insect, would somehow disrupt the balance, and the habitat would begin to deteriorate. Likewise, to add a species may be equally disruptive.

One of these pioneer ecologists was Frederick Clements, who studied ecology extensively throughout the Midwest and other areas in North America. He held that within any given region of climate, ecological communities tended to slowly converge toward a single endpoint, which he called the "climatic climax." This "climax" community was, in Clements's mind, the most well-balanced, integrated grouping of species that could occur within that particular region. Clements even thought that the process of ecological succession—the replacement of some species by others over time—was somewhat akin to the development of an organism, from embryo to adult. Clements thought that succession represented discrete stages in the development of the community (rather like infancy, childhood, and adolescence), terminating in the climatic "adult" stage, when the community became self-reproducing and succession ceased. Clements's view of the ecological community reflected the notion of a precise balance of nature.

Clements was challenged by another pioneer ecologist. Henry Gleason, who took the opposite view. Gleason viewed the community as largely a group of species with similar tolerances to the stresses imposed by climate and other factors typical of the region. Gleason saw the element of chance as important in influencing where species occurred. His concept of the community suggests that nature is not highly integrated. Gleason thought succession could take numerous directions, depending upon local circumstances.

Who was right? Many ecologists have made precise measurements, designed to test the assumptions of both the Clements and Gleason models. For instance, along mountain slopes, does one life zone, or habitat type, grade sharply or gradually into another? If the divisions are sharp, perhaps the reason is that the community is so well integrated, so holistic, so like Clements viewed it, that whole clusters of species must remain together. If the divisions are gradual, perhaps, as Gleason suggested, each species is responding individually to its environment, and clusters of species are not so integrated that they must always occur together.

It now appears that Gleason was far closer to the truth than Clements. The ecological community is largely an accidental assemblage of species with similar responses to a particular climate. Green

ash trees are found in association with plains cottonwood trees because both can survive well on floodplains and the competition between them is not so strong that only one can persevere. One ecological community often flows into another so gradually that it is next to impossible to say where one leaves off and the other begins. Communities are individualistic.

This is not to say that precise harmonies are not present within communities. Most flowering plants could not exist were it not for their pollinators—and vice versa. Predators, disease organisms, and competitors all influence the abundance and distribution of everything from oak trees to field mice. But if we see a precise balance of nature, it is largely an artifact of our perception, due to the illusion that nature, especially a complex system like a forest, seems so unchanging from one day to the next.

The discovery of the planet Pluto

Unlike Neptune, Pluto was discovered through a careful, systematic search, not simply by turning a telescope toward a position calculated on the basis of gravitational theory. Nevertheless, the history of the search for Pluto began with indications of deviations of the planets Uranus and Neptune from their predicted orbits. According to gravitational theory, such deviations from predicted orbit, or perturbations, would probably be caused by the gravitational pull of an unknown planet beyond the orbits of Uranus and Neptune, and the position and mass of that unknown planet could be calculated from the deviations. Early in the twentieth century, several astronomers became interested in this problem, including Percival Lowell, founder and director of Lowell Observatory in Arizona.

At the time Lowell made his calculations, Neptune had moved such a short distance since its discovery that it could not be used effectively to search for perturbations by an unknown planet. Therefore, Lowell and his contemporaries based their calculations primarily on minute irregularities in the motion of Uranus. Lowell's computations indicated two places where a perturbing planet could be, the more likely of the two being in the constellation of Gemini. He predicted a mass for the planet intermediate between that of Earth and that of Neptune (his calculations gave the predicted planet a mass of about 6.6 times the mass of Earth). Other astronomers, however, obtained other solutions, including one that indicated two unknown planets.

At his Arizona observatory, Lowell searched for the unknown planet from 1906 until his death in 1916, without success. Subsequently,

Lowell's brother donated to the observatory a photographic telescope that could record a 12-degree-by-14-degree area of the sky on a single photograph. The new camera went into operation in 1929, and the search was continued for the ninth planet.

Unfortunately, Gemini lies near the Milky Way (Earth's galaxy), and some 300,000 star images were recorded on each exposure. It was an immense task to compare all the star images on each of two or more photographs of the same field in the hope of finding one image that changed position with respect to the rest, revealing itself as the new planet. The job was facilitated by the invention of the blink microscope, a device for comparing two different photographs of the same region of the sky. The operator's vision automatically alternates between corresponding parts of the two photographs. If the star patterns are the same on the two plates, the observer sees a constant, although flickering, picture. However, if one object has moved slightly in the interval between the times the two plates were taken, the image of that object appears to jump back and forth as the view shifts between the two plates. In this way, moving objects can quickly be picked out from among thousands of star images.

In February 1930, Clyde Tombaugh, comparing photographs made on January 23 and 29 of that year, found an object whose motion appeared to be about right for a planet far beyond the orbit of Neptune. It was within 6 degrees of the position Lowell predicted for the unknown planet. The new planet was named Pluto, the god of the underworld. (Appropriately, the first two letters of Pluto are the initials of Percival Lowell; this is about as close as one can come to naming a planet for a person.)

Although in 1930 the discovery of Pluto appeared to be a vindication of gravitational theory similar to the nineteenth-century discovery of Neptune, we now know that Lowell's calculations were wrong. When the mass of Pluto was finally measured, it was found to be much less than that of the Moon. Such a small mass could not possibly have exerted any measurable pull on either Uranus or Neptune. Recently the Pioneer and Voyager spacecraft have penetrated beyond the orbit of Pluto, and they show no drift that might be attributed to an undiscovered mass. Further, a survey of the entire sky carried out in 1983 by the Infrared Astronomical Satellite revealed no hidden "Planet X." Today it is generally accepted that the supposed perturbations of Uranus and Neptune are not, and never were, real.