TESOL QUARTERLY Vol. 26, No. 4, Winter 1992

Demystifying the TOEFL® Reading Test

BONNY NORTON PEIRCE
Ontario Institute for Studies in Education

Despite the growing international influence of the TOEFL (Test of English as a Foreign Language), no articles have been published on how the test is actually developed by the Educational Testing Service (ETS). In this article, the author, who worked in the Test Development department at ETS from 1984 to 1987, seeks to demystify the TOEFL reading test at both a descriptive and theoretical level. First, the author draws on data from a case study of a reading test she developed in 1986 to illustrate the technical rigor with which the test is developed, and to raise questions about its theoretical adequacy. Second, the author draws on the theory of genre proposed by Kress (1989, 1991) to (a) illustrate how the unequal relationship between test makers and test takers predisposes TOEFL candidates to a particular reading of TOEFL texts; and (b) locate the TOEFL reading test within the larger social context of the TOEFL internationally, where competence in English means access to power. The author concludes that the TOEFL–2000 test development team at ETS, who are currently reviewing the test, needs to address the washback effect of the test in consultation with both ESOL teachers and TOEFL candidates internationally.

The global influence of the TOEFL (Test of English as a Foreign Language) is increasing at a rapid rate. In the 1988–1989 administration year, 566,000 candidates registered to take the TOEFL; in 1989–1990 this figure jumped to 675,000, climbing again in 1990–1991 to 741,000 (Educational Testing Service [ETS], 1990, 1991a, 1992). To date, however, there are no published articles on the way the TOEFL is actually created by test developers at the Educational Testing Service (ETS) in Princeton, New Jersey, U.S.A. Current publications on the TOEFL describe its history (Spolsky, 1990); compare it to other major ESOL tests (Bachman, Vanniarajan, & Lynch, 1988); or present the results of a wide variety of research questions on, among

®TOEFL is a registered trademark of Educational Testing Service.

others, the context bias of the TOEFL (Angoff, 1989), the TOEFL from a communicative point of view (Duran, Canale, Penfield, Stansfield, & Liskin-Gasparro, 1985), and TOEFL examinee characteristics (Wilson, 1982). However, as Raimes (1992) has suggested, research of this nature has not succeeded in demystifying the TOEFL for many TOEFL candidates and TESOL professionals, nor has it addressed more basic assumptions about what the TOEFL actually tests, why, and how. Given the fact that the TOEFL is currently undergoing review as part of the TOEFL–2000 project (Chyn, DeVincenzi, Ross, & Webster, 1992; ETS, 1991), a close examination of TOEFL test development procedures is timely.

In this paper, I wish to demystify the TOEFL reading test at a descriptive and theoretical level by drawing on my practical experience in the Test Development department at ETS from 1984 to 1987 and by drawing on theoretical insights from recent research in genre analysis. I will begin the paper with a brief description of the TOEFL as a whole, introduce some basic terminology used in psychometric testing, and describe the procedures I followed in the development of the TOEFL reading test. Thereafter, I will use a passage I assembled, reviewed, and pretested for a particular TOEFL reading test to illustrate how these test development procedures are put into practice and how the statistical analysis of individual items is used in the test development process. Thereafter, I will critically examine some of the assumptions I brought to the test development of the reading test, focusing on notions of authenticity, background knowledge, and test validity. I will then present the argument that a standardized reading test is best understood as a specific genre which presupposes an unequal relationship between test makers and test takers within the context of larger and frequently inequitable social structures. In this view, such a relationship has a significant impact on the social meaning of texts constituted within this genre. I will use these insights from genre analysis to help explain my case study data as well as to locate the TOEFL reading test within the larger social context of the TOEFL internationally. I will frame my concluding remarks with reference to possible innovations in TOEFL test development.

THE TOEFL

The TOEFL, first developed in 1963, is used to assess the English proficiency of candidates whose native language is not English, and scores are used by more than 2,400 colleges and universities in the United States and Canada to determine whether a candidate's level of proficiency in English is acceptable for the institution in question; the TOEFL is also used by institutions in other countries where English is

the medium of instruction (ETS, 1992). ETS does not determine passing or failing scores; the decision on which students are accepted by a particular institution is dependent on the policy makers of each individual institution. Policy varies from institution to institution, often depending on the kind of program a student has applied for and whether the institution offers supplementary courses in English. The test is administered by ETS on a monthly basis in approximately 1,250 test centers in 170 countries around the world (ETS, 1992). Any given TOEFL form is used only once, and the Test Development staff at ETS produces a new TOEFL form for each monthly administration. The test itself has a multiple-choice format and is divided into three sections: Section I, Listening Comprehension; Section 2, Structure and Written Expression; Section 3, Vocabulary and Reading Comprehension. The TOEFL Test of Written English, a short essay test, is included in five TOEFL administrations a year. The TOEFL Policy Council, comprising a Committee of Examiners, a Research Committee, and a Services Committee are responsible for different areas of program activity.

A short description of the pretesting process in the TOEFL helps to explain how one form of the TOEFL is made equivalent to another form. All TOEFL questions (items) are pretested on a sample TOEFL population. The experimental or pretest items are inserted into what is called the final form of a TOEFL. The final form contains all the items that have already gone through the pretesting process and been approved for use in a TOEFL administration. TOEFL candidates are tested on the inserted pretest items in the same way that they are tested on the final form items. (Candidates do not know which items are being pretested.) The pretest items are scored alongside the final form items, but the results on the pretest items are not calculated into the sample population's final TOEFL score. An item analysis is then conducted on each of the pretested items. The purpose of the item analysis is to determine the level of difficulty of each item as well as its discriminating ability; it also helps to pinpoint potential problems with the item. When the pretested items are ultimately used in a TOEFL final form, the statistics from the pretested items help a test developer assemble a final form with a level of difficulty equivalent to that of previous TOEFL tests. Items that have little discriminating ability are revised or discarded.

THE LANGUAGE OF MULTIPLE-CHOICE ITEMS

During the course of this paper, I will be using a number of psychometric terms and wish to introduce this vocabulary at the outset of the discussion. Consider the following question:

The answer in a multiple-choice question is referred to as
(A) an item
(B) a distractor
(C) an option
(D) a key

Choices A, B, C, and D are all referred to as *options*. The answer to the question, D in this case, is referred to as the *key*. The incorrect answers, A, B, and C in this case, are referred to as *distractors*. The question itself (*The answer to a multiple-choice question is referred to as*) is called the *stem*. The stem plus options are collectively referred to as the *item*. A reading comprehension passage combined with a number of items is referred to as a *set*.

THE TOEFL READING COMPREHENSION SECTION

Because the data on which I base my discussion is drawn from the reading section of the TOEFL, it is necessary to describe this section of the test in some detail. In Section 3 of the TOEFL, the Vocabulary and Reading Comprehension section, there are 30 vocabulary items and 30 reading comprehension items, and candidates are given 45 min to complete the section. When pretested items are included in the test, the number of items in the section increases to 90, and the time limits are modified. While all the vocabulary questions are discrete (individual) items, the reading comprehension section is divided into about five reading comprehension passages with approximately six items per reading passage. The five passages span a variety of disciplines—from passages with a focus in the humanities to passages with a more scientific focus. The TOEFL *Bulletin of Information for TOEFL/TWE and TSE, 1992–1993* (ETS, 1992) states that the Vocabulary and Reading Comprehension section of the test "measures ability to understand nontechnical reading matter" (p. 3) in standard written English. Candidates are required to answer the multiple-choice questions on the basis of what is "stated" or "implied" in each of the passages (p. 21), and they must choose what they consider the *best* of the four options provided in each item. The construct of reading that is measured in the TOEFL reading test is not made explicit in the ETS literature.

The passages that are chosen for the reading comprehension section are expository texts that have been drawn from academic magazines, books, newspapers, and encyclopedias; they are not written specifically for the TOEFL. To preserve the original quality of the texts, test developers are discouraged from changing the author's words, although deletions are permitted. The rationale for such a policy is that TOEFL candidates should be exposed to what is called authentic

language used by a variety of writers and not a customized "TOEFL English." However, passages that are potentially offensive are excluded from the test. A potentially offensive passage (called sensitive at ETS) is difficult to define and, in fact, a subject of much debate amongst members of the TOEFL test development team. At ETS, topics on politics, religion, or sex were considered to be sensitive topics. One of the main criteria whereby a passage was judged to be sensitive was its potential to create unnecessary anxiety for some candidates which would in turn compromise their performance on the test. A passage on birth control or abortion, for example, might create emotional stress for some candidates and lead to poor test performance. Sexist language (such as the use of the generic *he*) might be both offensive and ambiguous. In addition, topics that deal with a country other than a North American country might be perceived as giving unfair advantage to candidates who have background knowledge from the country in question. For example, a topic on the coffee industry in Brazil might be perceived by candidates from other parts of the world as giving unfair advantage to South Americans. It might also cause confusion amongst those South Americans who have a different understanding of the coffee industry from that of the author of the given text.

THE TEST DEVELOPMENT PROCESS

Because a new TOEFL form has to be produced each month, ETS trains private individuals outside ETS to perform the first step of the test development process. These individuals, known as item writers, are given assignments to find a variety of passages of appropriate length and to develop approximately six or seven items based on each passage. Item writers are given detailed test specifications to facilitate this process. The completed assignments are forwarded to ETS where a member of the test development team takes responsibility for converting the passage and items into a publishable pretest set. This was one of the functions I performed in the Test Development department at ETS. In the following paragraphs I will describe the procedures I followed in the development of items for the TOEFL reading test. While many of these procedures are standard practice at ETS, they are not unique to that institution (see Madsen, 1983).

When I was given an item writer's submission to develop for pretesting purposes, I did not initially refer to the questions that had been developed by the item writer because I wanted to explore my own response to the text before being influenced by the questions that the item writer had submitted. It was only after I had completed my own

analysis of the text and created my own questions that I would refer to the item writer's submissions and proceed with the revision process.

First, I adopted the position of a "reader" rather than a test developer in the initial stages of test development. I asked myself whether the text was interesting and held my attention. Where my concentration lapsed or I found myself rereading a portion of the text for clarification, I recorded my observations. Where there were stylistic shifts in the text, interesting use of metaphor or analogy, inferences and comparisons, contradictions or ambiguities, I made a note. At this initial stage I did not refer to the test specifications. I found that if I tried to rework the text and test items to suit the test specifications, I lost my reader response to the text. It is perhaps gratuitous to state that TOEFL candidates don't know what the test specifications are. Thus I tried to preserve, for as long as possible, my initial responses to the text. This enabled me to detect interesting nuances in the text as well as ambiguities and potential difficulties.

Second, once I had responded to the text as a reader, I began to examine the text as a test developer. One of the assumptions I made as I developed the items is that a reader's understanding of the text does not terminate once the reader has read the passage once or twice. I assumed that the longer readers work with a text, the more comprehensive their understanding of the text becomes. Indeed, a candidate's understanding of the test questions themselves also draws on the candidate's reading ability. Thus, with respect to the assessment of reading ability, there is an artificial distinction between the reading passage and the items, and I was sensitive to the fact that the test questions are an integral part of the process of reading comprehension. The main principles I followed when developing the items are summarized below.

Use the Candidates' Time Efficiently

Given the fact that the candidates have only 45 min for Section 3 of the TOEFL, which includes a vocabulary and reading section, and consequently less than 30 min for the reading section, one of my primary concerns as a test developer was to ensure that I used the candidates' time efficiently. I tried to ensure that there were as many items in a set as the passage could sustain; content that added little to the coherence of the passage and was not used for testing purposes was deleted from the text. I assumed it would be frustrating for candidates to grapple with portions of text and then have little opportunity to demonstrate their comprehension of this content. Furthermore, I tried to use closed rather than open stems. If the stem is open, that is, if there is no question mark at the end of the stem to consolidate the

thought expressed in the stem, candidates have to repeat the stem each time an option is read in order to follow the grammatical logic of the option. This is time-consuming and a burden on memory.

Help the Candidates Orient Themselves to the Text

Because all TOEFL reading passages have been removed from one context and transplanted in another context, I tried to help the candidates orient themselves to the content of the passages being tested. Because TOEFL passages do not have titles, I tried to ensure that the first item in a reading comprehension set addressed the main idea or subject matter of the passage. I hoped this would give candidates an organizing principle to help them in their attempts to develop a better understanding of more detailed aspects of the text. I was aware, however, that many texts do not have a main idea at all—they may be descriptive or narrative with little coherent argument as such. In such cases, a stem like *What does the passage mainly discuss?* would be preferable to *What is the main idea of the passage?* Furthermore, I assumed it would be generally helpful to candidates if the order of items in the set followed the order of information in the text itself. This would enable candidates to locate tested information with relative ease and enable them to build on their understanding of "old" or "given" information in the text. I used line references as much as possible, provided that they did not compromise the intent of the item (e.g., a scanning item). I was aware, however, that items which address the prevalent tone of the passage cannot always be directly associated with a particular line or sentence in a passage and are best left to the end of the item set.

Make Sure the Items Are Defensible

There are two issues that pertain to the defensibility of items: the items in combination and the items individually. With reference to the items in combination, I tried to ensure that the items did justice to the content and level of difficulty of the text. This is where the art of test development was central to the test development process. I had to use judgment and imagination to assess interesting (and uninteresting) characteristics of the passage and develop items that gave the candidates sufficient opportunity to demonstrate their understanding of these characteristics. For example, if the text was detailed and complex, I did not wish to underestimate the candidates' reading ability by asking trivial questions. In addition, I knew the same information from the passage could not be tested twice—albeit in different forms. To do so would place some candidates in double jeopardy. Conversely, the stem in one item could not reveal the answer to a question in another item.

For example, if the key to a particular item was *Gold is expensive,* then another item in the same set could not be worded, *Gold is expensive because . . . .*

With reference to individual items, I had to ensure that the stem and key of each item were unambiguous. Each stem had to contain as much information as was necessary and sufficient to answer the question posed, and each item had to have only one correct key. In addition, the options in any one item could not logically overlap in meaning. For example, if the key to an item were *The region suffered severe drought* and one of the distractors in the item were *The climatic conditions were harsh,* there would be logical overlap between the two options as the key would be encompassed within the distractor. This would create ambiguity and confusion for the candidates and present the candidates with two potential keys. Nevertheless, the distractors in an item needed to have some link to information in the text and they had to be plausible. Implausible distractors would be eliminated by candidates and lead them to choose the correct key by default. The distractors, however, could not be keyable. By this I mean that the distractors could not, potentially, be correct. This is an area that caused considerable debate in the test development process. Was a distractor drawing candidates because it was a good distractor or because it was ambiguous or potentially keyable? Furthermore, the item could not be keyable without reference to the passage; that is, it could not be keyable with reference to general background knowledge. Finally, no one option could stand out as being structurally or stylistically different from the other options. Thus if I put a definite article *the* in front of *key* in the example given earlier, Option D would stand out as different from the other options, which are preceded by indefinite articles. This could attract undue attention from candidates, who might (correctly) key it by default.

THE REVIEW PROCESS

Because it is not possible for one person to offer a definitive reading of a text or avoid all the potential problems associated with test development, a comprehensive review process has been developed at ETS. There are two cornerstones of the review process: first, a series of test reviews by approximately six different test development specialists; second, a pretesting process as described earlier in this paper. After the test developer is satisfied that the pretest has been adequately prepared, the test goes for a test specialist review (TSR). The test specialist reviewer (also TSR) is a member of the TOEFL test development team; indeed, all test developers are reviewers and all reviewers are test developers. The passage and items are systematically reviewed

by the TSR, who is simultaneously "taking" the test and reviewing it. The reviewer notes comments in a memo and returns it to the test developer. The test developer then works through these comments, makes appropriate changes to the items, and then arranges a meeting to discuss the review with the TSR. The test developer has to defend the action taken with respect to the TSR's suggestions.

After the test has gone through the TSR stage, it goes to the TOEFL coordinator who examines all the items again, two editors who focus on stylistic problems in the test, and a sensitivity reviewer who seeks to eliminate any potentially offensive material in the test. At each of these stages, the test is returned to the test developer for discussion and revision. During this entire process, the history of each item can be located by any one reviewer because all the reviews are kept in a folder until the test is ready for publication. Once galleys of the test have been made, it is then returned to the Test Development department for a final review before it is published in a TOEFL test booklet.

Each member of the team had her or his own style of reviewing. When I reviewed the test of another test developer, I had to make a decision about those items that I thought were acceptable, those that were definitely not acceptable, and those that had minor flaws. Thus reviewing a test could be quite a delicate process. On the one hand, I felt a responsibility to help create as defensible a test as possible; on the other hand, I didn't want to be unreasonable as this would compromise my efforts to defend those comments I felt strongly about. I was particularly concerned about items that I thought were ambiguous, had more than one potential key, or perhaps no clear key at all. I felt less strongly about items that were stylistically weak or had implausible distractors. If a test developer and reviewer could not resolve a problem that each felt strongly about, the issue was referred to a more senior member of the team for arbitration.

A TOEFL READING COMPREHENSION CASE STUDY

In order to illustrate the debates that arose in the test development process, and to contextualize the comments that I will make in the latter part of this paper, I will draw on the experience I went through as I developed one particular reading comprehension pretest for the TOEFL in 1986. There is nothing special about the passage and the items; they were chosen at random from a number of passages that I had developed, in collaboration with my colleagues, and which had gone through the pretesting and statistical analysis stages. If I had not chosen to use the passage and items for case study purposes, they would have been revised for the last stage of the test development

process: the final form. When I worked on the passage and the items, I had not anticipated that they would be used for the purposes of this exposition. Fortunately, however, I was able to locate the history of all the reviews in the ETS archives.

The passage that I have chosen to illustrate the above discussion is one that examines the farming of corn in the Middle West United States. Space does not permit a full discussion of all the items in the TSR and later reviews (though these can be found in Peirce, in press). I have chosen two items for the purposes of illustration and discussion because each raises a number of interesting theoretical issues about the nature of standardized reading tests.

The Passage

Running a farm in the Middle West today is likely to be a very expensive operation. This is particularly true in the Corn Belt, where the corn that fattens the bulk of the country's livestock is grown. The heart of the Corn Belt is in Iowa, Illinois, and Indiana, and it spreads into the neighboring states as well. The soil is extremely fertile, the rainfall is abundant and well distributed among the seasons, and there is a long, warm growing season. All this makes the land extremely valuable, twice as valuable, in fact, as the average farmland in the United States. When one adds to the cost of the land the cost of livestock, seed, buildings, machinery, fuel, and fertilizer, farming becomes a very expensive operation. Therefore many farmers are tenants and much of the land is owned by banks, insurance companies, or wealthy business people. These owners rent the land out to farmers, who generally provide machinery and labor. Some farms operate on contract to milling companies or meat-packing houses. Some large farms are actually owned by these industries. The companies buy up farms, put in managers to run them, provide the machinery to farm them, and take the produce for their own use. Machinery is often equipped with electric lighting to permit round-the-clock operation.

In general, all the reviewers found the passage to be acceptable for the purposes of the TOEFL. The only minor change took place when *businessmen* was changed to the current *business people* (Line 12) in keeping with a policy that encourages nonsexist language. The TSR did raise the following two issues, and then proceeded to resolve them:

1. Line 1—Do we need to say Middle West U.S.? Guess U.S. is in line 9 . . . Really seems like there should be a paragraph cut-off somewhere, but I guess that's authentic language for you.

In the interests of efficiency, the comments that are made by the various reviewers at ETS are written in an abbreviated style. Each test developer soon develops a style that is accessible to other members of the team. A common abbreviation however is *S:* This is short for *I*

suggest you do the following. In the first comment, the TSR is concerned that the introduction to the text does not offer sufficient geographical context for the reader, but is then satisfied that *United States* is mentioned elsewhere in the text. The second comment indicates that the reviewer would like to have the paragraph divided up for easier reading, but acknowledges TOEFL policy that discourages editorial changes in the interests of authenticity.

The Items

Note that the items are numbered in the 70s because they have been inserted into the final form of a TOEFL Vocabulary and Reading Comprehension section (normally 60 items) and numbered accordingly. The two items I am discussing are those that were presented to the TSR and the TOEFL coordinator. The revised items that were finally published in pretest form during an administration of the TOEFL are given in the Statistical Analysis section below. The complete set of pretested items is given in the Appendix.

Item 71 assesses a candidate's understanding of information that is not given explicitly in the passage but is strongly implied by the author in Lines 2 and 3 of the text.

71. It can be inferred from the passage that in the United States corn is
    (A) the least expensive food available
    (B) used primarily as animal feed
    (C) cut only at night
    (D) used to treat certain illnesses

Phrases used to introduce this item type would include *It can be inferred from the passage that; The passage supports which of the following conclusions?; The author implies that.* Comment 2 below was written by the TSR, Comment 3 by the coordinator.

2. 71. B—I think this only refers to Middle West corn. S: ". . . that Middle West corn is"—(A) only option that doesn't start w. verb. S: Sold at very low prices (for (B) could say "grown" instead of "used")

3. 71. C + D where do they come from?

The issues referred to above relate respectively to the accuracy of the stem, the stylistic quality of the options, and the suitability of the distractors. The first comment indicated that the use of *United States* was too vague, and I accordingly changed the stem to *Middle West.* The second comment indicated that Option A was not stylistically parallel to the other options. As I examined this option more carefully, I became conscious of two other problems with the option. The first was

that the word *expensive* had been used in the previous item—as the key—and was repeated in Item 75—again as the key. This overlap was undesirable as it might have attracted undue attention to these options. In addition, the use of *the least* made the distractor somewhat implausible: While it is plausible that corn might be an inexpensive product, it is far less likely to be *the least expensive food available.* I was therefore happy to use the reviewer's suggested revision. The TSR's final comment was written in parentheses to indicate that she did not feel strongly about the comment. Nevertheless, I was happy to change *used* to *grown* since the verb *used* was already present in Option D.

The coordinator's query (*C + D where do they come from?*) was an abbreviated way of asking how these particular distractors could be seen as plausible. I could defend Option C on the grounds that machines were used *round-the-clock.* It did occur to me however, that the key to another item (Item 78) was to be revised to *at night* and because I wanted to avoid overlap with this item, I proceeded to revise Option C to *cut in the morning.* I thought Option D was justified because of the repeated references to the word *operation* (Lines 2, 10, 18) which has a medical connotation. (In retrospect, however, I think the option is a weak one.) It was only after I had checked the statistics that came back from the pretesting of this item set that I realized there was a far more serious flaw in this item than any of the reviews had picked up. This will be discussed in the Statistical Analysis section below.

Item 75 calls for an understanding of information that is given explicitly in the passage:

75. According to the passage, a plot of farmland in an area outside the Corn Belt as compared to a plot of land inside the Corn Belt would probably be
    (A) less expensive
    (B) smaller
    (C) more fertile
    (D) more profitable

The answer to the question is clearly indicated in Lines 7–8 of the passage, which states that the land inside the Corn Belt is *extremely valuable, twice as valuable in fact, as the average farmland in the United States.* The first comment below was written by the TSR, the second by the coordinator.

4. 75. A.—(D) inferable, since the land would presumably cost less, hence less overhead S: easier (or "more mechanized"?) I wonder if (B) isn't inferable, too? S: less tiring (?)

5. 75. stem very wordy—any way to simplify?

Significantly, the TSR was as concerned with what could be reasonably inferred from the passage as with what was explicitly stated in the passage. Thus she argued that the key could not be defended simply on the basis of the opening phrase *According to the passage.* This is generally considered a last line of defense, but is best avoided in the interests of clarity and test quality. As I reflected on the TSR's comments, I could see why Option D might be construed as inferable and hence confusing to candidates. I took the suggestion that I should change the wording to *more mechanized.* In one of the later reviews, however, one of the editors took exception to the use of *more mechanized,* saying that it was "implausible" to call a plot of farmland mechanized. I changed the wording again to *more desirable.* At the time, I did not agree that Option B could be inferable. Logically, I believed that since the land outside the Corn Belt was depicted as less valuable—and hence less expensive—than that inside the Corn Belt, it was likely that the plots of farmland outside the Corn Belt would be larger and not smaller than those inside the Corn Belt. I took the position that the distractor was a good one, rather than an unfair one. I decided, however, that if another reviewer had a similar problem with Option B, I would change it. In the final analysis, a statistical analysis would tell me if Option B had presented problems to otherwise competent readers. In response to the coordinator's comment, I did simplify the stem without compromising the clarity of the question. The item that was finally pretested is presented in the Statistical Analysis section below.

Statistical Analysis

Once the passage and items had passed through all the reviews and I had adjusted the items where I thought necessary, the test was pretested in a TOEFL administration (see the Appendix). The results of the pretests were forwarded to the Statistical Analysis department at ETS, which completed an item analysis on each item and forwarded the results to the Test Development department. The work of the test developers at this stage was to assess the results of the item analyses, decide which items worked, which needed to be revised, and which needed to be discarded. How does a test developer know whether an item has "worked"? In standardized reading tests a successful item is one that discriminates successfully between "good" and "poor" candidates. The level of difficulty of an item, on the other hand, is a function of the percentage of candidates who choose the correct key. The latter statistic is not difficult to compute. However, the test developer needs to be assured that the relative difficulty of the item is a function of the relative levels of proficiency of the candidates—as measured by the test—and not a function of a poorly constructed or ambiguous item.

In order to determine whether an item discriminates successfully between good and poor candidates, there needs to be a criterion (standard) by which to judge the item. The criterion that is used in the TOEFL reading test is the candidates' performance on Section 3 of the TOEFL. Thus, for example, an item is considered to have "worked" if most of the top candidates in the Vocabulary and Reading Comprehension section get the item right and if candidates who choose the correct key are not randomly distributed through the sample. If the latter were the case, the item would have no discriminating power. In order to determine who the top candidates are, each candidate's total score on Section 3 is computed, and candidates are given percentile rankings. On the basis of these percentile rankings, the total group is then divided into 5 subgroups, ranging from the top 20% to the bottom 20%. Once this information is tabulated, the performance of each individual item is determined with respect to these 5 groups. The index of discrimination, the biserial correlation, is a correlation coefficient that measures the extent to which candidates who score high on Section 3 as a whole tend to get the item right, and those who score low tend to get it wrong. The item is working successfully if the biserial correlation is above .5.

In the passage that I had pretested, all the items except one can be considered to have discriminated successfully between the candidates. The biserial correlations for all the items except Item 71 were above .5, and the item set was judged to be of average difficulty for the TOEFL population. What then was the problem with Item 71, which had in fact been revised considerably (see below)? In the population on which my reading comprehension passage was pretested there were 1,280 candidates, all of whom were divided into five different groups of 256 candidates based on percentile rankings. (See Table 1, a simplified form of an ETS item analysis.) Note that by the time Item 71 was pretested, 9 candidates in the two weakest groups had dropped out, which explains the slight discrepancy in the Total row at the bottom of Table 1. A candidate who has "dropped out" is no longer attempting

TABLE 1
Item 71

Percentile rank	0–20%	21–40%	41–60%	61–80%	81–100%	Total
Omit	4	0	5	4	2	15
A	85	105	109	103	57	459
B (Key)	95	110	120	143	196	664
C	27	15	5	4	1	52
D	37	25	17	2	0	81
Total	248	255	256	256	256	1271

to answer any questions; a candidate who "omits" an item is still nevertheless attempting to answer all questions and is therefore included in the Total figures. The item analysis follows.

71. It can be inferred from the passage that Middle West corn is
    (A) sold at very low prices
    (B) grown primarily as animal feed
    (C) cut in the morning
    (D) used to treat certain illnesses

As a preliminary analysis of Item 71, compare the candidates who chose Option A, a distractor, with those who chose Option B, the key. A large number of candidates in the weakest group chose the distractor A as the key (85 in all), while a smaller group (57) in the strongest group chose the distractor A as the key. Significantly, the situation is reversed for Option B, the key: While only 95 of the weakest candidates correctly chose Option B as the key, 196 of the strongest candidates correctly chose Option B as the key. A cursory glance indicates that the item is working quite well: 52% of the candidates chose the correct option—664 out of a total of 1,271 candidates—a moderately difficult item. It is clear that Option A was the most attractive distractor as 459 of the total 1,271 candidates chose this option as the key; 52 candidates chose Option C; 81 chose Option D.

Despite these apparently favorable results, there are some disturbing issues that arise from the analysis: An uncomfortably high number of candidates chose Option A as the key—160 of whom were in the top two groups—and 15 candidates omitted this item—6 of whom were in the top two groups. It was for these reasons that the biserial correlation fell below .5 to .35, and I carefully scrutinized the item. As I reexamined the key in Item 71 and the information in the passage from which it was drawn, it became clear to me that the key, strictly speaking, was inaccurate. The passage states that most of the livestock in the United States is fed on corn that originates in the Corn Belt. This does *not* imply, however, that Middle West corn is grown primarily as animal feed. Although this may indeed be the case in the United States, such an inference cannot be drawn from the passage per se. For example, the corn could be grown primarily for export purposes, even though it is the staple diet for livestock in the United States. In a TOEFL final form, the item would have needed to be revised or excluded from the item set.

By way of comparison, consider the statistical analysis of Item 75 (see Table 2). The item read as follows:

75. According to the passage, a plot of farmland in an area outside the Corn Belt as compared to one inside the Corn Belt would probably be

(A) less expensive
(B) smaller
(C) more fertile
(D) more desirable

TABLE 2
Item 75

Percentile rank	0–20%	21–40%	41–60%	61–80%	81–100%	Total
Omit	3	4	0	1	1	9
A (Key)	66	101	149	191	222	729
B	62	44	41	29	21	197
C	73	61	41	22	6	203
D	39	43	24	13	6	125
Total	243	253	255	256	256	1263

In Item 75, which had a biserial correlation of .55, only 66 candidates in the weakest group correctly chose Option A as the key, whereas 222 in the strongest group correctly chose Option A as the key. In contrast, 62 of the weakest group incorrectly chose Option B as the key, whereas only 21 in the strongest group incorrectly chose Option B as the key. Similar comparisons can be drawn with Options C and D. In total, 9 candidates omitted the item, only 2 of whom were in the top two groups. The percentage of candidates who chose the correct key was 58% (729 of a total 1,263)—thus the item was of average difficulty. The remaining candidates were relatively evenly divided in their choice of distractors. Nevertheless, it was still a little disturbing that 50 candidates in the top two groups incorrectly chose Option B as the key. The comment of my TSR reviewer had been validated. This distractor would have needed revision before it reached the final form stage.

DISCUSSION

I have demonstrated in the above discussion that TOEFL test development procedures incorporate a complex set of checks and balances which include both qualitative and quantitative feedback. With reference to qualitative feedback, I have demonstrated that the development of the TOEFL reading test is a collaborative effort in which test developers work with colleagues to minimize ambiguity and confusion within individual items. Such collaboration gives test developers the opportunity to subject TOEFL texts and items to alternative readings and interpretations. With reference to quantitative feedback, I have demonstrated that the statistical analysis of pretested items provides a

different kind of feedback for the test developer. It may confirm the reservations that the test developer has had about a particular item; it may draw attention to aspects of the item that have been overlooked; it may help to resolve disputes about the suitability of an item. However, notwithstanding the technical rigor with which the TOEFL reading test is developed, the above discussion raises a number of critical questions about the assumptions I brought to the test development process. The questions I wish to address concern the three related issues of authenticity, background knowledge, and test validity.

Authenticity

I have suggested that TOEFL test developers strive to utilize "authentic" reading passages: Passages used in the TOEFL are extracted from "real" texts, and test developers are discouraged from tampering with these extracts. As demonstrated in the discussion above, while the TSR wished to put in a paragraph break in the text on the United States corn industry, she resisted the desire to do so because of the policy on authentic language. If there were no paragraph break in the original text, she assumed the extract would have the same meaning as the original only if the paragraph break were omitted. I concurred with this observation.

In retrospect, however, it is clear that this approach to authenticity is flawed, both at the level of textual meaning and at the level of social meaning. First, at the level of textual meaning: If a passage is extracted from a larger text, and readers have no access to this larger text—the type of text, the title, the author, the intended audience, the date of publication, the publisher—the extract has little resemblance to its authentic textual origins. Furthermore, if parts of this extract are deleted for test development purposes, the extract has even less claim to authenticity. Second, at the level of social meaning: As argued by educators who adopt a poststructuralist approach to text (Belsey, 1980; Hill & Parry, 1992; Morgan, 1987; Peirce, 1991; Simon, 1992—to name a few), the meaning of a text is not only derived from what an author "demonstrates" in a text but also from the conditions under which the text is received. In poststructuralist theory, "meaning" therefore refers not only to the sentence-level meaning of a text but also to its social meaning. The social meaning of a text is constituted at the intersection between the words in the text, the reader's investment in the text, and the particular space/time location in which the text is read. In this view, using the author's words in a standardized reading test does not guarantee that the text will have the same meaning as the original from which it was extracted, notwithstanding attempts at au-

thenticity; its meaning derives from the interaction between the text, the test taker, and the testing situation in which the text is read.

By way of illustration, one need only examine the text I have used in this article. Under what conditions might I, as a student in the "real" world, be interested in the United States corn industry? If I were waiting for the campus doctor and picked up the text to pass the time, I would approach it from one perspective; if I needed to read the text for an oral presentation in a business course, I would approach it from another perspective; if I were taking the TOEFL exam, my approach, yet again, would be radically different. On each of these occasions, the value ascribed to the text—the social meaning of the text—would be different because the social conditions under which I was reading it and the purpose for which I was reading it would vary considerably: If I were passing time in a doctor's office, the points that I would find salient in the text would be mediated by my own personal interest in the topic and perhaps by a certain degree of anxiety about the condition of my health at that point in time. If I were reading the text for an oral presentation in a business course, the salient points in the text would be mediated by the questions that I brought to the text and my perception of what my fellow students and instructor would find interesting in the larger context of the business course. If I were reading the text for a TOEFL exam, the points that I would find salient would be mediated almost entirely by the questions that the test maker had formulated. The extent to which this unequal test maker/test taker relationship predisposes the TOEFL candidates to a particular reading of TOEFL texts is discussed further in the Genre Analysis section below.

Background Knowledge

The second question I wish to raise concerns the place of background knowledge in standardized reading tests. I raise this question because, although I have stated that TOEFL test developers strive not to test a candidate's background knowledge, a *TESOL Quarterly* reviewer claimed that s/he could answer Question 75 without reference to the text. While this issue deserves fuller attention (see, e.g., Clapham, 1991), I will address only two concerns that arise from the reviewer's observation. First, if a TOEFL candidate answers this question correctly (and the key is not randomly chosen) one of the following assumptions can be made: (a) The candidate (call him A) knows nothing about real estate prices in the United States but has read and understood *both the passage and the question.* (b) The candidate (call her B) has background knowledge about real estate prices in the United States and has read and understood *the question alone.* In other words, whether or not a

candidate has background knowledge about real estate prices in the United States, the candidate still has to have sufficient command of the English language to understand *the question* in order to answer it correctly. I made a similar point earlier when I argued that a distinction between the passage and the questions is an artificial one—both the passage and the questions test a candidate's reading ability. If the language of the question is easier than the language of the text, Candidate B would have an advantage over Candidate A with respect to background knowledge and time. However, if the language of the question is no simpler than the language of the text, then the only advantage that Candidate B would have over Candidate A would be a time advantage. That is, she would not have to take up valuable time to consult the text.

Second, while I do not wish to trivialize a time advantage in a test situation, I think there is a more fundamental issue at stake here which pertains to the nature of the testing situation. Hill and Parry (1992) have argued convincingly that in standardized reading tests, "personal knowledge must be continually suppressed for fear of making an inappropriate response" (p. 458). When candidates come to the test situation, they assume that the background knowledge they already have must not interfere with the knowledge that is "demonstrated" in the texts that they will be required to read: Candidate B's source of information about real estate prices might be different from that of the author, or the text might be out of date. It would not be in Candidate B's best interests to trust her own judgment and knowledge in order to answer Question 75. TOEFL test developers are well aware of this dilemma, which is why all questions of this nature are prefaced by phrases such as *According to the passage.* As Hill and Parry argue, it is indeed ironic that readers "are encouraged to hold separate the very knowledge which is crucial to their effective engagement with text" (p. 458).

Test Validity

Simply put, a valid test is one that measures what it intends to measure (Henning, 1987). The TOEFL claims to measure a candidate's ability to understand nontechnical reading matter. To what extent can this claim be upheld? I have argued that attempts at authenticity are flawed. I have argued that if a candidate has background knowledge of reading matter that is actually tested in a TOEFL reading test, it may not be in the candidate's best interests to use it. I have also demonstrated that the acceptability of an item is determined with reference to performance of the candidates on Section 3 as a whole. In other words, the quality of TOEFL reading test Item X is a function

of the quality of the sum total of all the items in the TOEFL section of which Item X is one part. The only conclusion that can confidently be drawn is that if a candidate performs well on the TOEFL reading test, the TOEFL candidate is a good reader of TOEFL tests. Thus when I judged the acceptability of the items I had pretested, and judged all of them, apart from Item 75, to successfully discriminate between "good" and "poor" TOEFL candidates, I did so on the basis of a self-referential criterion, rather than an independent measure of reading ability. This has implications for the validity of the TOEFL reading test. While the TOEFL can confidently claim to measure a candidate's ability to read nontechnical reading matter *in a TOEFL test*, the extent to which these measures apply to preparation for an undergraduate oral presentation, the doctor's waiting room, and the countless other occasions in which the candidate reads nontechnical reading matter must be called into question.

GENRE ANALYSIS AND THE TOEFL READING TEST

Having critically examined some of the assumptions I brought to the development of the TOEFL reading test, I wish to argue that the conception of a standardized reading test as a particular genre is a theoretically useful lens through which to examine my case study data as well as the location of the TOEFL reading test within the larger social context of the TOEFL internationally. While genre analysis has been utilized in a wide variety of fields such as literary studies, linguistics, and rhetoric (see Swales, 1990), it has not as yet made a significant impact on the field of language testing. Following Kress (1989, 1991), who draws on the poststructuralist theory of Foucault (1977), the concept of genre I wish to use in this paper is that of genre as "text"—either oral or written—constituted within and by a specific social occasion which has a conventionalized structure and which functions within the context of larger institutional and social processes.

In Kress's formulation, the social occasions which constitute a genre may be formulaic and ritualized, such as a wedding or committee meeting, or less ritualized, such as a casual conversation. The important point is that the conventionalized forms of these occasions and the organization, purpose, and intentions of social participants within the occasion give rise to the meanings associated with the specific genre, whether it be a tutorial, interview or—as I will argue—a standardized reading test. Furthermore, Kress (1989) has demonstrated that increasing difference in the power relations between participants in an interaction has a particular effect on the social meaning of the texts within a particular genre. In essence, in genres in which there is great power

difference between the social participants, the *mechanism* of interaction, the conventionalized form of the genre, is most foregrounded, while the *substance* of the interaction, the content, is least foregrounded. The conception of genre that Kress is proposing, which foregrounds the centrality of power within a particular social occasion in the context of larger social processes, is a departure from more conventional approaches to genre analysis. The latter tend to present genres as uncontroversial forms of texts such as sonnets, term papers, and interviews, with little reference to the larger and frequently inequitable social structures in which these texts are constituted.

While test makers have generally assumed that a standardized reading test is an aberration in the "real" world, I wish to argue that it is no less authentic a social situation than an oral presentation or a visit to a doctor. In a standardized reading test, the value ascribed to texts within this genre and the meaning that is constructed is associated with a ritualized social occasion in which participants share a common purpose and set of expectations. The social occasion is characterized by strict time limits in which test takers have little control over the rate of flow of information in the activity—what Peirce, Swain, and Hart (in press) refer to as the "locus of control" in the activity. The test takers are expected to be silent at all times, respect rigorous proctoring procedures, and read the text in solitude. As Hill and Parry (1992) argue, social behavior in a testing situation is tantamount to cheating. Both test makers and test takers recognize that the purpose of the test is to discriminate between readers of varying levels of proficiency with reference to a criterion established a priori by the test makers. The expectations are that the background knowledge of the test takers has little relevance to the items being tested and that the test makers decide what an acceptable reading of the text should be. Thus the relationship between the test makers and the test takers, a manifestly unequal one, has a direct bearing on the social meaning ascribed to texts in the standardized reading test. Furthermore, the standardized reading test must be understood with reference to larger social processes in which test takers have unequal access to material, educational, and linguistic resources: While some test takers have comfortable homes where literacy material is commonplace, superior educational opportunities, and familiarity with the conventions of standardized reading tests, other test takers have no electricity in their homes, limited access to literacy material, and few educational opportunities.

The conception of the standardized reading test as genre helps explain data from the case study described above. Consider for example the statistical analysis of Item 71. Item 71 was a flawed item: There was no key. Nevertheless, over 66% of the candidates in the top two categories (339 of 512 candidates) chose the intended key. Apparently,

these candidates knew the conventions of standardized reading tests: One of the options was intended to be correct; their task was to determine which one of the four options I, as the test maker, had in mind. Although the words in the question were inappropriate, the test takers sought to understand what I meant, not what I said: In other words, how did I, as the test maker, intend the TOEFL candidates to "read" the text? The unequal relationship between me as a test maker and the candidates as test takers had a direct bearing on the social meaning of the text. The test takers' personal investment in the test mitigated against their objecting to the poor quality of the item. In such a social situation, they had to conform to the conventionalized rules of the test, or resist at great personal cost. It is significant that of the 512 candidates in the top two categories who examined this question, only 6 exercised the dubious right to omit responding to the flawed question.

Furthermore, the conception of the standardized reading test as genre leads to an examination of the location of the TOEFL reading test within the larger context of the TOEFL internationally. In this spirit, the point that needs to be stressed is that the TOEFL is not just any standardized test—it is the largest test of English in a world that has adopted English as its lingua franca. For this reason, people who are considered to have command of the English language not only have linguistic versatility but educational, economic, and political power (Peirce, 1989; Pennycook, 1992; Phillipson, 1992; Tollefson, 1991). A test that determines who has command of the English language has inordinate power to influence not only the educational future of individuals but the political future of nations. This is the larger social context of the TOEFL, the link between TOEFL test makers, TOEFL test takers, and larger social processes in which competence in English means access to power. Thus the social meaning of texts used in the TOEFL are constituted with reference to the test takers' personal investment in a uniquely powerful standardized test.

CONCLUSION: IMPLICATIONS FOR TOEFL–2000

Given the location of the TOEFL with respect to the increasing power of the English language internationally, the challenge for the TOEFL–2000 project at ETS is to determine whose interests the TOEFL serves and how the TOEFL can best serve those interests. The TOEFL–2000 committee has stated that the aim of the TOEFL is not simply to serve the interests of admissions officers at United States and Canadian universities. It has indicated that ETS is committed to serving the ESL/EFL community, and that it wishes to "better reflect current understanding of language and communication and second language

learning and testing" (ETS, 1991c). Given the expressed interest in serving the ESL/EFL community, I would like to suggest that revisions to the TOEFL be made with reference to a consideration of the "washback" effect of the TOEFL.

The washback effect of a test, sometimes referred to as the systemic validity of a test (Alderson & Wall, 1992), refers to the impact of a test on classroom pedagogy, curriculum development, and educational policy. Swain (1985) indicates that a concern with washback was a guiding principle in the development of communicative language tests for French immersion programs in Canada. Wesche (1987) states that interest in positive washback was of primary concern in the development of the Ontario Test of English as a Second Language. Recent research by Shohamy (1992) has found that the introduction of three national language tests in Israel has had a dramatic impact on classroom pedagogy and educational policy in the country. With reference to the TOEFL, however, reports on the washback effect of the test remain anecdotal. Its effect can only be extrapolated from the vibrant industry in TOEFL preparation books.

When the TOEFL is revised as part of the TOEFL–2000 project, TOEFL test developers and TOEFL consultants should take the opportunity to consider what kind of impact the TOEFL has had on classroom pedagogy and educational policy not only in North America but in some of the 170 countries in which the TOEFL is administered. ESL teachers internationally should be consulted to determine what construct of reading should be assessed in the TOEFL reading test and how the test can best serve the interests of their programs and the needs of their students. TOEFL candidates should be consulted to determine how preparation for the TOEFL could promote language learning as well as improved test-taking strategies, and how the anxiety associated with the TOEFL testing situation could be alleviated. The outcome of such research has the potential to transform the current relationship between TOEFL test makers and TOEFL test takers—between those who "have" and "have not" command of the English language internationally.

ACKNOWLEDGMENTS

I would like to acknowledge former ETS colleagues in the Languages Group of the Test Development department, who contributed in many ways to the development of this paper. I would also like to thank Merrill Swain, Kathleen Troy, Kate Parry, David Mendelsohn, and Sandra Silberstein for their insightful comments on an earlier draft of the paper; two *TESOL Quarterly* reviewers for their rigorous critiques; ETS for giving me access to TOEFL archives; and the Social Sciences and Humanities Research Council of Canada for its financial support.

THE AUTHOR

Bonny N. Peirce, a PhD candidate in the Modern Language Centre, Ontario Institute for Studies in Education, Canada, is interested in the relationship between social theory and the practical concerns raised by second language learners, teachers, and testers internationally. She is a corecipient of the 1990 Malkemes Prize for her 1989 *TESOL Quarterly* article, "Toward a Pedagogy of Possibility in the Teaching of English Internationally: People's English in South Africa."

REFERENCES

Alderson, J. C., & Wall, D. (1992). *Does washback exist?* Paper presented at the 14th Annual Language Testing Research Colloquium, Vancouver, Canada.

Angoff, W. (1989). *Context bias in the Test of English as a Foreign Language* (TOEFL Research Rep. No. 29). Princeton, NJ: Educational Testing Service.

Bachman, L., Vanniarajan, A. K. S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing, 5,* 128–159.

Belsey, C. (1980). *Critical practice.* London: Methuen.

Chyn, S., DeVincenzi, F., Ross, J., & Webster, R. (1992). *TOEFL–2000: Update.* Paper presented at the 26th Annual TESOL Convention, Vancouver, Canada.

Clapham, C. (1991). *The effect of academic discipline on reading test performance.* Paper presented at the 13th Annual Language Testing Research Colloquium, Princeton, NJ.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C., & Liskin-Gasparro, J. E. (1985). *TOEFL from a communicative viewpoint of language proficiency* (TOEFL Research Rep. No. 17). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1990). *Bulletin of information for TOEFL and TSE, 1990–91.* Princeton NJ: Author.

Educational Testing Service. (1991a). *Bulletin of Information for TOEFL and TSE, 1991–92.* Princeton NJ: Author.

Educational Testing Service. (1991b, Spring). *Newsline.* Princeton, NJ: Author.

Educational Testing Service. (1991c). *TOEFL–2000: Planning for change.* Princeton, NJ: Author.

Educational Testing Service. (1992). *Bulletin of information for TOEFL/TWE and TSE, 1992–93.* Princeton, NJ: Author.

Foucault, M. (1977). What is an author? In D. Bouchard (Ed.), *Language, countermemory, practice.* Ithaca, NY: Cornell University Press.

Henning, G. (1987). *A guide to language testing.* Cambridge, MA: Newbury House.

Hill, C., & Parry, K. (1992). The test at the gate: Models of literacy. *TESOL Quarterly, 24*(3), 433–461.

Kress, G. R. (1989). *Linguistic processes in sociocultural practice.* Oxford: Oxford University Press.

Kress, G. R. (1991). Critical discourse analysis. *Annual Review of Applied Linguistics, 1990, 11,* 84–99.

Madsen, H. (1983). *Techniques in testing.* New York: Oxford University Press.

Morgan, R. (1987). Three dreams of language. *College English, 49,* 449–458.

Peirce, B. N. (1989). Toward a pedagogy of possibility in the teaching of English internationally: People's English in South Africa. *TESOL Quarterly, 23*(3), 401–420.

Peirce, B. N. (1991). Review of the TOEFL Test of Written English (TWE) Scoring Guide. *TESOL Quarterly, 25*(1), 159–163.
Peirce, B. N. (in press), The development of a TOEFL reading test. In C. Hill & K. Parry (Eds.), *Testing and assessment: International perspectives on English literacy.* Harlow, England: Longman.
Peirce, B. N., Swain, M., & Hart, D. (in press). Self-assessment, French immersion, and locus of control. *Applied Linguistics.*
Pennycook, A. (1992). *The cultural politics of teaching English in the world.* Unpublished doctoral dissertation. Ontario Institute for Studies in Education/University of Toronto.
Phillipson, R. (1992). *Linguistic imperialism.* Oxford: Oxford University Press.
Raimes, A. (1992). Comments on "The TOEFL Test of Written English: Causes for concern". The author responds to Traugott, Dunkel, and Carrell. *TESOL Quarterly, 26*(1), 186–190.
Simon, R. I. (1992). *Beyond the racist text. Teaching against the grain: Texts for a pedagogy of possibility.* Toronto: OISE Press.
Spolsky, B. (1990). The prehistory of TOEFL. *Language Testing, 7,* 98–118.
Shohamy, E. (1992). *The power of tests: A study on the impact of language tests on teaching and learning.* Paper presented at the 14th Annual Language Testing Research Colloquium, Vancouver, Canada.
Swain, M. (1985). Large-scale communicative language testing: A case study. In S. Savignon & M. Burns (Eds.), *Initiatives in communicative language teaching.* Reading, MA: Addison-Wesley.
Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.
Tollefson, J. W. (1991). *Planning language, planning inequality.* New York: Longman.
Wesche, M. (1987) Second language performance testing: The Ontario Test of ESL as an example. *Language Testing, 4,* 28–47.
Wilson, K. (1982). *A comparative analysis of TOEFL examinee characteristics, 1977–1979* (TOEFL Research Rep. No. 11). Princeton, NJ: Educational Testing Service.

APPENDIX

Questions 70–78

Running a farm in the Middle West today is likely to be a very expensive operation. This is particularly true in the Corn Belt, where the corn that fattens the bulk of the country's livestock is grown. The heart of the Corn Belt is in Iowa, Illinois, and Indiana, and it spreads into the neighboring states as well. The soil is extremely fertile, the rainfall is abundant and well-distributed among the seasons, and there is a long, warm growing season. All this makes the land extremely valuable, twice as valuable, in fact, as the average farmland in the United States. When one adds to the cost of the land the cost of livestock, seed, buildings, machinery, fuel, and fertilizer, farming becomes a very expensive operation. Therefore many farmers are tenants and much of the land is owned by banks, insurance companies, or wealthy business people. These owners rent the land out to farmers, who generally provide machinery and labor. Some farms operate on contract to milling companies or meat-packing houses. Some large farms are actually owned by these industries. The companies buy up farms, put in managers to run them, provide the machinery to farm them, and take the produce for their own use. Machinery is often equipped with electric lighting to permit round-the-clock operation.

70. What is the author's main point?

(A) It is difficult to raise cattle.
(B) Machinery is essential to today's farming.
(C) Corn can grow only in certain climates.
(D) It is expensive to farm in the Middle West.

71. It can be inferred from the passage that Middle West corn is

(A) sold at very low prices
(B) grown primarily as animal feed
(C) cut in the morning
(D) used to treat certain illnesses

72. In line 3, the word "heart" could best be replaced by which of the following?

(A) spirit
(B) courage
(C) cause
(D) center

73. It can be inferred from the passage that the region known as the Corn Belt is so named because it

(A) is shaped like an ear of corn
(B) resembles a long yellow belt
(C) grows most of the nation's corn
(D) provides the livestock hides for leather belts

74. The author mentions all of the following as features of the Corn Belt EXCEPT

(A) rich soil
(B) warm weather
(C) cheap labor
(D) plentiful rainfall

75. According to the passage, a plot of farmland in an area outside the Corn Belt as compared to one inside the Corn Belt would probably be

(A) less expensive
(B) smaller
(C) more fertile
(D) more desirable

76. As described in the passage, which of the following is most clearly analogous to the relationship between insurance company and tenant farmer?

(A) Doctor and patient
(B) Factory owner and worker
(C) Manufacturer
(D) Business executive and secretary

77. The word "their" in line 15 refers to

(A) companies
(B) farms
(C) managers
(D) machinery

78. According to the passage, some machinery is equipped with electric lighting so that it can be used

(A) indoors
(B) in the fog
(C) at night
(D) while it rains

Note. From the Test of English as a Foreign Language (Form 3IATF10), 1986. Princeton, NJ: Educational Testing Service. Copyright 1986 by Educational Testing Service. Reprinted by permission of Educational Testing Service.