2 The Test of English as a Foreign Language: developing items for reading comprehension

Bonny Norton Peirce

The purpose of this chapter is two-fold. First, I wish to share the insights that I gained about test development while working in the Languages Group of the Test Development department at the Educational Testing Service (ETS) from 1984 to 1987. Second, I wish to describe the creation of a particular reading comprehension test I prepared for a TOEFL (Test of English as a Foreign Language) administration in 1986. I will begin the chapter with a brief description of the TOEFL as a whole, introduce readers to some basic terminology used in psychometric testing, and discuss some of the procedures I followed in the development of a TOEFL reading comprehension test. Thereafter, I shall use a passage I assembled, reviewed, and pre-tested for a TOEFL reading test to illustrate some of the debates that arise in the process of test development and to show the kind of information that is provided by a statistical analysis of individual items in a reading comprehension test. Additional analysis of the passage and items can be found in Peirce (1992).

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 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 test is administered by ETS on a monthly basis in more than 1250 test centres in 170 countries around the world (ETS 1992). The number of candidates taking the TOEFL test is increasing rapidly. In the 1988–89 administration year, 566 000 candidates registered to take the TOEFL; in 1989–90 this figure jumped to 675 000 – climbing again in 1990–91 to 740 000 (ETS 1990, 1991, 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 programme a student has applied for and whether the institution offers supplementary courses in English. Any given TOEFL form is used only once, and the Test Development staff at ETS produce a new TOEFL form for each monthly administration. The test itself has a multiple-choice format and is divided into three sections: Section 1, 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, is responsible for different areas of programme activity.

A short description of the pre-testing process in the TOEFL helps to explain how one form of the TOEFL is made equivalent to another form. All TOEFL questions (items) are pre-tested on a sample TOEFL population. The experimental or 'pre-test' items are inserted into what is called the 'final form' of a TOEFL test. The final form contains all the items that have already gone through the pre-testing process and been approved for use in a TOEFL administration. TOEFL candidates are tested on the inserted pre-test items in the same way that they are tested on the final form items (candidates do not know which items are being pre-tested). The pre-test items are scored alongside the final form items, but the results on the pre-test items are not calculated into the sample population's final TOEFL score. An item analysis is then conducted on each of the pre-tested 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 pre-tested items are ultimately used in a TOEFL final form, the statistics from the pre-tested 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

A number of psychometric terms will be used in the course of this chapter. The following question illustrates many of these terms:

The answer in a multiple-choice question is referred to as

(A) an item
(B) a distractor
(C) an option
(D) a key

(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 in this chapter is drawn from the reading section of the TOEFL,[1] 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 minutes to complete the section. When pre-tested 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. Candidates are required to answer the multiple-choice questions on the basis of what is 'stated' or 'implied' in each of the passages and they must choose what they consider the best of the four options provided in each item. The *Bulletin of Information for TOEFL/TWE and TSE 1992-1993* (ETS 1992, p. 3) states that the Vocabulary and Reading Comprehension section of the test 'measures ability to understand nontechnical reading matter' in standard written English. The construct of reading that is measured in the TOEFL reading test is not made explicit.

The passages that are chosen for the reading comprehension section are expository texts that have been drawn from academic magazines, books, newspapers, and encyclopaedias; they are not written specifically for the TOEFL. Test developers are discouraged from changing the words used by the author, 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. Passages that are potentially offensive are excluded from the test. A potentially offensive passage (called 'sensitive' at ETS) is difficult to define, and is in fact a subject of much debate amongst members of the TOEFL test development team. At ETS, topics on politics, religion, or sex are considered to be sensitive. One of the main criteria whereby a passage is judged to be sensitive is 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 than 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 pre-test set. This was one of the functions I performed in the Test Development department at ETS. In the following paragraphs I shall 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 this institution (see Madsen 1983).

When I was given an item writer's submission to develop for pre-testing purposes, I did not initially refer to the questions that had been developed by the item writer because I wished 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 re-reading 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 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 summarised below.

1. Use the candidates' time efficiently

Given the fact that the candidates have only 45 minutes for Section 3 of the TOEFL, which includes a vocabulary and a reading section, and consequently less than 30 minutes 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.

2. 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 organising 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 generally be 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, providing 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.

3. 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 judgement 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. This 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 among the candidates and present them 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? 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 above, (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; and, second, a pre-testing process, as described earlier in this chapter. After the test developer is satisfied that the pre-test 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 simultaneously 'takes' the test and reviews it. The reviewer notes down all the comments on a memo and returns them to the test developer. The test developer then works through these comments, makes changes to the items where he or she thinks them appropriate, and then sets up a meeting to discuss the review with the TSR. The test developer has to defend the action that he or she has taken with respect to the TSR's suggestions.

Each member of the team has his or her 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.

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 tries 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 publishing. Once galleys of the test have been made, it is returned to Test Development for a final review before being published in a TOEFL test booklet.

A TOEFL reading comprehension case study

In order to illustrate the debates that arise in the test development process, and to contextualise the comments that I will make in the latter part of this chapter, I shall draw on the experience I went through as I developed one particular reading comprehension pre-test 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 pre-testing 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 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 United States of America.

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 non-sexist language. The TSR did raise the following two issues, and then proceeded to resolve them:

line 1 – Do we need to say 'Middle West U.S.?' Guess U.S. is in line 8

– 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 on 'authentic language' that discourages editorial changes in the interests of authenticity.

The items

Note that the items are numbered from 70 to 78 because they have been inserted into the final form of a TOEFL reading comprehension section (normally 60 items) and numbered accordingly. The items I am discussing are those that were presented to the Test Specialist Reviewer. The items that were finally published in pre-test form during a 'live' administration of the TOEFL are given in the appendix to this chapter.

The first item in the set is a question that assesses a candidate's overall understanding of the main topic of the passage.

70. What is the author's main point?

(A) Livestock are expensive to raise.
(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.

The key (D) is drawn primarily from the topic sentence in the first line of the passage: 'Running a farm in the Middle West today is likely to be a very expensive operation'. This idea is also supported by the comment in lines 10-11, '...farming becomes a very expensive operation'. The comments below were written by the TSR and coordinator respectively.

70. D – neat question!
70. (A) livestock is?

While the TSR is satisfied with the item as it stands, the coordinator raises the question of agreement between the subject of the sentence in option (A) 'livestock' and the verb 'are'. It would have been simple for me to change 'livestock are' to 'livestock is', but when I took another look at the item, it became apparent to me that the word 'expensive' was used twice in the item – in both options (A) and (D). As (D) is the key, this is particularly problematic. It seemed to me that the repetition of the word 'expensive' might attract attention to these two options and candidates might key (D) by default. I therefore chose to revise option (A) completely to read 'It is difficult to raise cattle'. An alternative format could have been used to test the same information. Consider:

Which of the following would be the best title for the passage?

(A) Raising Cattle: Problems and Solutions
(B) The History of Farming: A Changing Landscape
(C) Growing Corn: The Role of Climate
(D) Farming in the Middle West: Money Matters

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

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

This kind of item does not rely on a candidate's background knowledge, but is drawn from information that is given explicitly in the text. Phrases used to introduce this item type include: 'It can be inferred from the passage that...', 'The passage supports which of the following conclusions?', 'The author implies that...'. The comments below were written by the TSR and coordinator respectively.

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')

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 (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. In addition, I was aware that the use of such exclusive terms as *the least* and *only* are often perceived by candidates to be distractors rather than keys because of the unlikelihood of such extremes occurring at any one particular time. In other words, 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 brackets 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 last comment ('where do (C) and (D) come from?') was an abbreviated way of asking how these particular distractors could be seen as plausible. I could defend (C) on the grounds that machines are used 'round-the-clock', and I thought (D) was justified because of the repeated references to the word 'operation' (lines 2, 11, 18) which has a medical connotation. 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). As a result of all these deliberations, which in real time would take no more than a few minutes, I changed the item to read:

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

It was only once I had checked the statistics that came back from the pre-testing of this item set that I realised there was a far more serious flaw in this item than any of the reviews had picked up (this will be discussed under 'Statistical analysis' below).

Item 72 assesses a candidate's understanding of particular words and phrases as they are used in the context of the passage.

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

(A) Spirit
(B) Courage
(C) Cause
(D) Center

This type of item is distinct from items in the vocabulary section of the TOEFL in that all the options refer to possible synonyms of the word 'heart' in different contexts. Thus for (A) one might say that a person has a compassionate heart/spirit; for (B) that a person should take heart/courage; for (C) that the heart/cause of a problem is that ... ; for (D) (the key) that the heart/center of the Corn Belt is in ... The suggestion below, which was simple to implement, was made by the TSR.

72. D —these options need to be lower case, since they're replacing 'heart'

Item 73 tests a reader's ability to understand the use of metaphor in the text. The TSR's comment on the item is given below.

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

73. C – nice yet humorous

On reflection, it is apparent that this item is not an inference that can be drawn directly from the passage. Rather, it tests information that can only be extrapolated from the passage. In a final form, I would revise this item to read: 'It is likely that the region known as the Corn Belt is so named because it ...'

Item 74 assesses a candidate's understanding of information that is given explicitly in the passage.

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

(A) rich soil
(B) advantageous weather
(C) cheap labor
(D) sufficient rainfall

The form of this item is irregular because the question is phrased negatively. In order for candidates to answer this question correctly, they cannot simply focus on one option – they have to examine all four options carefully and arrive at the key by elimination. Although the stem indicates that the author 'mentions' certain features of the Corn Belt in the text, the options do not contain information taken verbatim from the text as this would make the item far too easy for the candidate population. Thus 'rich soil' must be equated with 'extremely fertile soil' (line 4), 'advantageous weather' with 'long, warm growing season' (lines 5/6) and so forth. The comments below were written by the TSR and coordinator respectively.

74. C. – (B), (D) similar – (D) is encompassed in (B). S: (B) warm weather

74. (D) in psg the rainfall is 'abundant'

The TSR drew my attention to the fact that I had collapsed options (B) and (D), thus making the item a 3-option item and reducing the attractiveness of both options. The TSR's suggested revision 'warm weather' made option (B) sufficiently distinct from (D) and I accordingly made the change. The coordinator drew my attention to the fact that 'sufficient' is not synonymous with 'abundant' and therefore might be construed as keyable. I therefore changed this option to 'plentiful' rainfall.

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 comments below were written by the TSR and coordinator respectively.

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

75. stem very wordy – any way to simplify?

Significantly, the reviewers were as concerned with what could be reasonably inferred from the passage as with what was explicitly stated in the passage. Thus they 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 defence, but is best avoided in the interests of clarity and test fairness.

As I reflected on their comments, I could see why (D) might be construed as inferrable 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 mechanised. I changed the wording again to 'more desirable'. At the time, I did not agree that (B) could be inferrable. 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 (B) I would change the option. In the final analysis, I knew that a statistical analysis would tell me if (B) had presented problems to otherwise competent readers. In response to the third comment, I did simplify the stem without compromising the clarity of the question.

Item 76 presents material that is analogous to material that is presented in the passage and asks candidates to recognise this relationship.

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 and retailer
(D) Business executive and secretary

It is significant that this type of item does not draw on information that is either explicitly stated in the passage or directly inferrable from the passage. It asks candidates to recognise a particular relationship that is described in the text, and it does so by presenting a variety of relationships to candidates – only one of which is analogous to the relationship described in the text. Thus the candidates are required to extrapolate from ideas that are given in the text. The only comment on this item was written by the TSR.

76. B – hope vocab doesn't get in the way here (esp. 'retailer')

The TSR expressed concern that the language in the item was more complex than the language in the text. Thus, even if the candidates could understand the relationship between the insurance company and tenant farmer, they would not know the equivalent relationship expressed in the options. I could easily revise 'retailer' to 'merchant' but I could not avoid using the term 'analogy' without compromising the item as a whole. I was also uncomfortably aware that the relationships depicted in the options might be unfamiliar to some candidates, though I was confident that the vast majority of candidates would be familiar with an owner/worker relationship. I decided to wait for a statistical analysis before abandoning the item at the pre-test stage.

Item 77 assesses the reader's understanding of the way cohesive devices are used in the passage to link intrasentential relationships: '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.' The only comment on this item was made by the TSR.

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

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

77. A – (D) only singular. S: machines

All the options are taken directly from the text since the item does not test knowledge of vocabulary but knowledge of syntax. I was therefore loath to follow the reviewer's suggestion to change 'machinery' to a plural form because the use of a synonym rather than the word used in the text might confuse the candidates and compromise the intent of the item.

Item 78 began its history as a straightforward item that assessed a candidate's ability to understand information that is given explicitly in the passage. The only comment in this item was made by the TSR.

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) twenty-four hours a day
(D) while it rains

78. C – key quite a bit longer. S: (C) at night.

The TSR's comment on this item indicates concern that the candidates might choose (C) as the key because its length attracts attention and not because the candidates have understood the reference 'round-the-clock'. I therefore took the suggestion to change the option to 'at night'. On reflection however, although this revision improves the item stylistically, it makes the key weaker. The passage states explicitly that machinery is equipped with electric lighting so that it can be used 'round-the-clock'. The phrase 'twenty-four hours a day' is thus a stronger key than the phrase 'at night', which is really an inference that is drawn from the passage and not a statement that is made explicitly. However, if I had turned this question into an inference question by saying, 'It can be inferred from the passage that...' I think that all the options would have been keyable. A better revision might have been to leave the key as 'twenty-four hours a day' and then make at least one of the distractors a little longer to balance the length of the key. For example, I could have changed (D) to 'when it rains unexpectedly'.

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 pre-tested in a TOEFL administration (see appendix). The results of the pre-tests were forwarded to the Statistics Department at ETS who completed an analysis of each item and forwarded the results to Test Development. The task 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 a test like the TOEFL, test takers, teachers, test developers, and administrators are particularly concerned with two issues: first, that the TOEFL discriminates successfully between 'good' and 'poor' candidates; second, that one form of the TOEFL is comparable in difficulty with other forms of the TOEFL. For test development purposes, 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 chose 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 five sub-groups, ranging from the top 20 per cent to the bottom 20%. Once this information is tabulated, the performance of each individual item is determined with respect to these five groups. The index of discrimination, the biserial correlation, is a correlation coefficient that measures the extent to which candidates who scored high on Section 3 as a whole tended to get the item right, and those who scored low tended to get it wrong. The item is working successfully if the biserial correlation is above 0.5.

In the passage that I had pre-tested, all the items except one can be considered to have been successful. All the biserial correlations except item 71 were above 0.5, and the item set was judged to be of average difficulty for the TOEFL population. From easiest to most difficult, the items ranked as follows: 72, 74, 78, 70, 77, 76, 73, 75, 71. What, then, was the problem with item 71, which had in fact been revised considerably? In the population on which my reading comprehension passage was pre-tested there were 1280 candidates, all of whom were divided into five different groups of 256 candidates based on percentile rankings. (See Table 2.1, a simplified form of an ETS item analysis.) Note that by the time item 71 was pre-tested, 9 candidates in the two weakest groups had dropped out, which explains the slight discrepancy in the 'Total' row at the bottom of Table 2.1. A candidate who has 'dropped out' is no longer attempting 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.

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 number (57) in the strongest group chose the distractor (A) as the key. Significantly, the situation is reversed for (B), the key: while only 95 of the weakest candidates correctly chose (B) as the key, 196 of the strongest candidates correctly chose (B) as the key. A cursory glance indicates that the item is working quite well: 52 per cent of the candidates chose the correct option: 664 out of a total of 1271 candidates – a moderately difficult item. It is clear that option (A) was the most attractive distractor as 459 of the total 1271 candidates chose this option as the key; 52 candidates chose option (C); 81 chose option (D).

Despite these apparently favourable results, there are some disturbing issues that arise from the analysis: an uncomfortably high number of candidates chose (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

Table 2.1 Responses to 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

Table 2.2 Responses to 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	256	256	256	1263

below 0.5 to 0.35 and I carefully scrutinised the item. As I re-examined 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 need to be revised or excluded from the item set.

By way of comparison, consider the statistical analysis of item 75 (see Table 2.2). The item reads 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

In item 75, which had a biserial correlation of 0.55, only 66 candidates in the weakest group correctly chose (A) as the key, while 222 in the strongest group correctly chose (A) as the key. In contrast, 62 of the weakest group incorrectly chose (B) as the key, while only 21 in the strongest group incorrectly chose (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 per cent (729 of a total 1263) – 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 (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.

Conclusion

The above discussion has highlighted a number of important issues in the development of the TOEFL reading test at ETS. First, test development is always a collaborative effort in which test developers work with colleagues to enhance the quality of the reading test. Such collaboration gives test developers the opportunity to subject the passage and items to alternative readings and minimise ambiguity in individual items. Second, the pre-testing process and the statistical analysis of pre-tested items provides a different set of checks and balances 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 fairness and suitability of the item. However, despite the rigour with which the TOEFL reading test is developed, important questions have been raised both within and outside ETS about the validity of the TOEFL reading test. While it is not within the scope of this chapter to address these questions, the reader may refer to Peirce (1992) for an analysis of some of the debates. As Robert Altman, the vice-president of ETS, argued in his plenary address at the International TESOL convention in Vancouver in 1992, machine-scorable standardised tests may not adequately reflect what students really know.

Acknowledgements

I would like to acknowledge the members of the Languages Group, Test Development, ETS, for the diverse ways in which they have contributed to the production of this chapter; in particular, Barbara Suomi, Valerie Richardson, and Ellie LeBaron for their part in the case study. I would also like to thank Susan Chyn and Jackie Ross for their comments on an earlier draft of the chapter, and ETS for providing access to TOEFL archives and TOEFL copyright material.

Note

1. Large sections of this chapter were drawn from the author's article published in *TESOL Quarterly*, 1992, 23:4, p. 665-691, 'Demystifying the TOEFL reading test'. The author wishes to thank *TESOL Quarterly* for permission to reprint these sections.

Appendix 2.1: A TOEFL pre-test item

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 and merchant
(D) Business executive and secretary

77. The word 'their' in line 17 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 the Eductional Testing Service. Reprinted by permission of the ETS TOEFL Program Office.

3 An entrance test to Japanese universities: social and historical context

John E. Ingulsrud

Historical development

Civil service examinations in ancient China

Modern life has been so greatly influenced by western civilisation that for many people 'modernisation' is virtually synonymous with 'westernisation'; and yet, certain features fundamental to modernity come not from the west but from the east. Consider, for example, institutional bureaucracy. It was in East Asian societies that such bureaucracy first arose out of the notion that the persons who lead should be the most able. The selection of leaders, it was thought, should be based on individual worth, not on endowed wealth or status. Of course, such meritocratic ideals existed outside East Asia as well. Plato, for example, in *The Republic* describes how society should choose its 'guardians':

> ...so we must introduce our Guardians when they are young to fear and, by contrast, give them opportunities for pleasure, proving them far more rigorously than we prove gold in the furnace...And any Guardian who survives these continuous trials in childhood, youth, and manhood unscathed, shall be given authority in our state. (Plato 1987, p. 180)

Through these 'trials' individuals were to be selected for their superior attributes of mind and body.

It was only in the Chinese empire of the Han Dynasty (206BC – AD 220), however, that these ideals were actually applied on a nationwide scale. The selection process for the bureaucracy was carried out by means of tests – not of physical abilities such as archery or javelin throwing – but of abilities demonstrated through the skills of reading and writing. The civil service examinations for the imperial bureaucracy took the form of, from our perspective, an achievement test. They were based on a specific body of knowledge which consisted largely of the literary