IRRI RESEARCH PAPER SERIES
Number 105 February 1985

A READING AND LISTENING COMPREHENSION TEST IN ENGLISH FOR NONNATIVE SPEAKERS APPLYING FOR TRAINING AT IRRI

E. P. CERVANTES, D. R. MINNICK, P. S. GOSECO,
and R. MAMINTA

The International Rice Research Institute
P.O. Box 933, Manila, Philippines

A READING AND LISTENING COMPREHENSION TEST IN ENGLISH FOR NONNATIVE SPEAKERS APPLYING FOR TRAINING AT IRRI[1]

ABSTRACT

This study developed a proficiency test in reading and listening comprehension, to be used to screen nonnative speakers of English applying for training at the International Rice Research Institute. The study is based on the concept of teaching English for Special Purposes.

Developing the test involved seven steps: 1) writing the test, 2) reviewing the test items, 3) pretesting, 4) administering the test, 5) statistical analysis, 6) retaining and weeding out items, and 7) assembling the final form. The communicative needs of the participants and the test content were determined by analyzing printed and audio-tutorial materials used in IRRI's training programs, and by interviewing trainers.

Of the 376 items examined, 48% were at a desirable level of difficulty (neither too easy nor too difficult) and 85% were highly discriminating. Based on the difficulty and discrimination indices of each item, 291 items were retained in the final form. Forty per cent of the items met both ideal criteria: desirable level of difficulty and high discrimination. The reliability coefficient of the test was 0.99045, indicating that the items are highly intercorrelated.

Although results show the test is valid and reliable, there is a need to revalidate results. The test should be administered to a larger sample. It can also be given to nonagriculturists to see whether it measures the English proficiency of an individual or his knowledge of agriculture. Other methods of estimating the reliability of the test should be used. A study of the correlation between an examinee's score on the English test and his later performance as an IRRI trainee should be conducted.

ERRATA for IRPS 105

Page 9, Table 4.

Column 6, 4th entry: 55 should be 35.
Last column, 4th entry: 29 should be 23.

Page 9, paragraph 1, line 2

"The stem . . ." should read "The item . . ."

[1] By Emerita P. Cervantes, editorial assistant, Communication and Publications Department, and Dan R. Minnick, head, Training and Technology Transfer Department, IRRI; Petronila S. Goseco, associate professor, and Rosario Maminta, chairman, Language Teaching Department, University of the Philippines System. Submitted to the IRRI Research Paper Series Committee October 1984.

A READING AND LISTENING COMPREHENSION TEST IN ENGLISH FOR NONNATIVE SPEAKERS APPLYING FOR TRAINING AT IRRI

The International Rice Research Institute (IRRI), an institution devoted to increasing the productivity of rice lands in developing countries, trains about 300 researchers and extension workers every year. Trainees study rice culture and rice-based cropping systems, and learn improved crop production techniques. The training period usually varies from 2 wk to 6 mo.

Most trainees come from non-English countries such as Bangladesh, Burma, China, Cuba, India, Indonesia, Germany, Japan, Korea, Malaysia, Pakistan, Philippines, Thailand, and Vietnam.

An IRRI trainee spends about 50% of his time listening to lectures, 25% reading handouts and other reading materials, and 25% doing field work. At free time, he can use self-instructional materials consisting of tapes and scripts available at the Training and Technology Transfer Department. In all these activities, English is the medium of instruction. A trainee lacking the necessary vocabulary and reading and listening proficiency finds himself at a disadvantage, gaining little from the training.

The study sought to formulate a proficiency test in reading (including vocabulary) and listening comprehension in English, which would be used to screen nonnative speakers of English applying for training at IRRI. Specifically, the study sought answers to:

1. How is a proficiency test in reading and listening comprehension for agricultural-training applicants developed?
2. Will the resulting test be valid and reliable? How do we test validity and reliability?

REVIEW OF RELATED STUDIES

In the United States, several English-language proficiency tests are widely used to screen nonnative speakers of English seeking admission to colleges and universities.

The Test of English as a Foreign Language (TOEFL), administered by the Educational Testing Service, is the most widely used English-language proficiency test. TOEFL allows nonnative speakers of English to demonstrate their English proficiency at the advanced level required for study in American institutions of higher learning. The American Language Institute, Georgetown University (ALIGU), has prepared and published a battery of tests for measuring the English proficiency of applicants for training grants, scholarships, and fellowships made available to foreign nationals by U. S. government agencies. The English Language Institute of the University of Hawaii has developed an English-language proficiency test for class placement purposes. The test measures one's ability to read with understanding and reasonable speed, vocabulary proficiency, ability to understand normal spoken English, and facility in written self-expression.

In Thailand, Sunkatup (13) developed an integrative/pragmatic English proficiency test for fourth year student teachers. The test consisted of aural discourse, lecturette, dictation, dictocomp, reading comprehension, maze, error recognition, and cloze. Sunkatup found tha of the skills measured, writing and listening were the most difficult. He cited a need to teach and evaluate language skills using the holistic, integrative approach. He recommended that the test be standardized after including a test of speaking skills and other types of integrative tests (those that test two or more points together, as contrasted with discrete-point tests which test only one point).

In the Philippines, many persons have developed English-language proficiency tests for various groups.

Villar (15) developed an integrative proficiency test in English that would measure the reading, writing, and listening skills of fourth year high school students of Nueva Viscaya. The test consisted of a reading comprehension test, cloze, dictation, maze, listening comprehension, dictocomp, and lecturette. The close, reading comprehension, maze, and listening comprehension tests proved to be the most valid and reliable devices for Villar's sample. She suggested that the test be used to gather feedback on the effectiveness of the language program in the secondary schools, especially on the teaching-learning materials, teaching methods, and their evaluation. She recommended standardizing the test and correlating it with other standardized tests such as the National College Entrance Examination, achievement tests, and creativity tests.

Manilay (9) constructed an English proficiency test for six groups of Asian students. The test was limited to grammar and constructed around the common linguistic problems of Thais, Pakistanis, Indonesians, Chinese, Japanese, and Vietnamese. These linguistic problems were revealed through already existing contrastive analyses of English and the native languages of the group, and through

interviews with native speakers of those Asian languages. The test was linguistically valid and highly reliable. Manilay recommended that the test be used with members of these Asian groups holding bachelor's degrees in English, and that it be administered to a larger sample. She also recommended more teaching of English grammar to improve English proficiency.

Francisco (5) and Cuevas (3), working with the same Asian groups, developed a test on vocabulary and a test on pronunciation. The vocabulary items were based on vocabulary problems of the second language speakers as pointed out by previous contrastive analyses (comparing the native languages with English) and interviews. Both tests were linguistically valid and highly discriminating.

Nartea (10) developed and standardized a 260-item English proficiency test for fourth year high school students of agricultural schools. The testing objectives were based on instructional objectives achieved by the target group. The test consisted of discrete-point and integrative types of test. The discrete-point items measured phonology, vocabulary, and structure. The integrative components tested reading, listening, and writing skills. The test had an acceptable difficulty level (0.43) and a high coefficient of correlation (0.91 by the Pearson Product Moment Formula, and 0.81 by the Kuder-Richardson Formula 20). The concurrent validity of the test was 0.47 and the predictive validity was 0.56. Nartea recommended using the test in agricultural schools in Visayas and Mindanao with populations similar to those she studied. She also recommended that concurrent validity be further determined by correlating examinees' scores with their scores in another standardized test. She suggested that a test of speaking skills be developed and evaluated.

All these tests have been constructed around linguistic problems determined through contrastive analysis of the participants' native language and English, a syllabus or course of study in English, or instructional objectives achieved by the target group.

English for Special Purposes or English for Specific Purposes (ESP) implies that English be taught and tested *according to the specific needs of the participant*, determined by an analysis of his needs. An ESP test does not test *General English*. ESP requires a clear definition of the purpose for which the participant will use English. This study developed an ESP test.

CONCEPTUAL FRAMEWORK

Certain changes and developments in education, especially in foreign language teaching, have produced a climate in which English is needed for various purposes. The changes have influenced the way English is taught so as to enable the learner to apply this language to his particular needs. The shift from *teacher-centered instruction* to *learner-centered instruction* has led to a search for individualization and specialization of instruction. It demands that the teaching be designed to meet the precise needs of the learner (12).

English, once taught only as an element in a general education, began to be taught for practical communicative uses. English began to be offered at different age levels for specific needs.

ESP is usually defined with reference to an occupational requirement or to professional study. ESP emphasizes the *purpose* of the learner for learning the language. This purpose determines the precise area of language required, the skills needed, and the range of functions to which the language will be put. English is taught not as an end in itself but as an essential means to a clearly identifiable goal (8).

In the primary and secondary levels of education where English is generally taught as a foreign language, students rarely have an immediate and specific requirement for using the language in any communicative situation. In ESP, however, the student has a special occupational or educational purpose. He may need to learn English because he has to communicate with English-speaking professors in a university or because he has to communicate with English-speaking tourists in the hotel where he will work.

Strevens (12) lists testing or assessment as one of four types of problems arising in ESP. He states that a major shortcoming of ESP is that very little work has been done to devise fresh methods of testing, examining, and assessment that match the new courses of training. ESP teachers have rejected as unsuitable virtually all of the existing tests in English as a Foreign Language. Alderson (1) states that perhaps the most wonderful argument for specific tests is that they produce a diagnostic profile which can match the student to the communicative needs of his particular course of study. It is necessary to specify the communicative needs of testees because of the inadequacy of previous test instruments; "there is always a number of students who have to abandon their studies and return home because of their language inadequacy and the progress of a much larger number is adversely affected in one way or another by language problems" (1).

Developers of ESP tests must distinguish between *usage* and *use*. If we accept Webster's definition of usage as "habitual and customary" and use to mean "toward an objective," this distinction implies that ESP should emphasize use, not usage. In other words, the test should not be based on items chosen on linguistic grounds alone, but more important, on how the testee requires to use the language. From the *use* point of view, language should be taught and tested according to the specific needs of the learner.

ESP tests also need *authenticity*. All tasks undertaken should be real-life, interactive communicative operations. Walsh (16) states that one rationale for using authentic text is that students will perceive the relevance of the task and will be more strongly motivated and better disposed toward the task. Morrow (1979, cited in 2) maintains that conventional tests seem to ignore certain features of genuine language use, which are: that language is interactive, purposive, authentic, contextualized, based in performance,

and assessed in terms of behavioral outcomes. Authenticity is often overlooked in the search for reliability and economy, which often results in a monotonous series of linguistic manipulations that are not closely related to real communicative tasks.

DESIGN AND METHODOLOGY

Designing the test

Identifying the participants. The participants are nonnative speakers of English applying for training at IRRI. They come from countries where English is a third or a second language, often taught as an optional course in college, or sometimes not at all. These countries include Bangladesh, Bhutan, Burma, China, Cuba, Germany, India, Indonesia, Japan, Korea, Malaysia, Nepal, Nigeria, Pakistan, Philippines, Thailand, and Vietnam. The participants have some knowledge of agriculture, acquired in school or from experience. They will communicate with lecturers and trainers who speak different English accents, listen to lectures and seminars, read handouts and other printed materials, use self-instructional materials such as slides, tapes, and scripts, all in English, and do field work under the supervision of English-speaking trainers.

Analyzing the participant's communicative needs. The participants' communicative needs were determined by textual analysis and by interviewing trainers. Lecture handouts used in the training programs, audio-tutorial modules, and some IRRI publications were analyzed to determine the vocabulary, microskills, functions, and linguistic components required of a trainee. Trainers were asked what English language skills they feel a trainee must have, to perform well in and benefit from the training course.

In the textual analysis, the kinds of passages were noted. For example, what does this passage do? Does it define a process, describe steps, compare methods, enumerate advantages, etc? What linguistic components were used to fulfill those functions? What connectors, intrasentential devices, modal auxiliaries, and other devices were used?

The textual analysis showed that the participant needs basic vocabulary for such functions as defining, identifying, enumerating, describing, explaining, exemplifying, comparing, contrasting, and generalizing.

Preparing the test content. Based on the textual analysis and interviews with trainers, test content was specified (see Table 1).

Developing the test

Test development involved seven steps: 1) writing the test, 2) reviewing the test items, 3) pretesting, 4) administering the test, 5) statistical analysis, 6) retaining and weeding out items, and 7) assembling the final form.

Writing the test. Materials for the test were drawn from lecture handouts, self-instruction modules, and IRRI publications. The lexical items chosen were those encountered most often and used not only in a specific field but across disciplines. Lexical items and linguistic components selected for testing were placed in a natural setting, ie, used in a sentence or passage reflecting the kind of materials encountered in the training program. Highly technical terms were avoided.

Tests were placed in proper sequence. The reading comprehension test proceeded from understanding lexical items to understanding relations between parts of a sentence, and finally to understanding relations between parts of a text. The listening comprehension test proceeded from understanding sentences to understanding short passages, and finally to understanding a lecture.

The test consisted of 376 items (283 in the Reading test and 93 in the Listening Comprehension), long enough to allow discarding of poor or nondiscriminating items (up to one-third of the total number of items). A brief description of the test follows:

Reading Test

I. Vocabulary. Examinee reads the sentence containing the lexical item being tested and selects from three options the meaning of the word as used.

II. Ordering Sentences. Sentences of a paragraph have been disarranged. Examinee chooses from four options the correct sentence order.

III. Maze. Examinee reads a 5-paragraph passage and predicts 25 words through 3 alternatives arranged vertically.

IV. Sentence Comprehension. Examinee completes a sentence by choosing from four options. The test calls for an understanding of relations within the sentence.

V. Reading Comprehension. Examinee reads the passages and answers questions that follow each. Questions are true-false and multiple choice.

VI. Cloze. Examinee reads 3 passages in which 55 words have been deleted. Examinee predicts those words.

Listening Comprehension Test

I. Sentence Comprehension. Examinee hears a problem sentence and chooses from three illustrations or tables the one that is described or that answers the problem.

II. Question Comprehension. Examinee hears a question. He chooses from four printed options the one that correctly answers the question.

III Aural Discourse. Examinee hears 10 short passages (3-6 sentences long) and answers true-false questions about each.

IV. Following Oral Directions. Examinee hears a series of instructions and does what he is instructed to do. All instructions require the examinee to indicate his answers in the answer sheet.

V. Lecture. This is similar to aural discourse. However, the material is longer and examinee is encouraged to take down notes during the lecture. A notes sheet which lists the points the examinee should take is

Table 1. Test content.[a]

	Microskill 1	Language realization	Mode
Reading and vocabulary	1. deducing the meaning of unfamiliar lexical items through contextual clues and roots		Sentences
	2. understanding relations within the sentence, especially intrasentential connectors and modal auxiliaries	exemplification, consequence contrast, comparison, addition, cause and effect, explanation, time, purpose, generalization, enumeration, defining and nondefining restrictive clauses, etc.	Sentences
	3. understanding relations between parts of a text through the grammatical cohesion devices of reference, substitution, ellipsis, and logical connectors 4. scanning to locate information 5. understanding explicitly stated information 6. understanding information not explicit 7. converting diagrammatic information into written statements		Maze Ordering sentences Cloze Reading passages
Listening Comprehension	1. recognizing vocabulary used in a sentence heard 2. recognizing key lexical items that signal a question		Sentences Pictures Questions
	3. identifying topic of a passage 4. following topic development 5. getting the important points of a passage		Aural discourse
	6. recognizing key words in oral instructions		Series of related instructions Lecture
	7. recognizing functions of intonation to signal information structure 8. selecting key words and using abbreviations and symbols 9. reconstructing notes into sentences		

[a]Has been modified in this paper to include only general descriptions. Details of language realization have been intentionally omitted to avoid giving away test content to prospective examinees.

provided. After the lecture, examinee answers multiple-choice questions about the lecture and summarizes what he has heard by reconstructing his notes. (The notes sheet is also collected to see if examinee took down notes during the lecture.)

Reviewing the test items. The test was placed aside for 2 wk and then reviewed by the writer. The test was later reviewed by two nonagriculturists proficient in English. The reviewers also took the test to see if it measures English proficiency or knowledge of agriculture. The results showed that a person with no background in agriculture but with English proficiency can do well in the test.

The materials for the Listening Comprehension Test were read by a female Tagalog speaker who has some background in agriculture, good diction, and a more-or-less neutral English accent. The materials were read at normal speed and recorded on a good-quality tape with a professional taping equipment. The tape was recorded without distortions and other distracting sounds.

Pretesting. A pilot run on 23 November 1983 enabled the writer to determine how much time to allow for the test, what aspects of test administration needed improvement, and whether the speaker for the taped Listening Comprehension test could be understood.

The examinees needed an average of 5 h to complete the test. This is understandable considering that most examinees had little or no English in college.

Comments on the taped test were solicited. For example, was the volume too loud or too soft? Were there mispronounced words? Was the speaker difficult to understand? Was the time allowed for each item too short? Materials were then retaped, considering all comments.

Administering the test. The trial run was conducted on 3 separate dates: 18, 26, and 30 November 1983. A total of 67 examinees took the test. A larger sample would have been more desirable because "Other things being equal, a larger sample is more likely to mirror the characteristic of the population than a smaller one" (7). Ingram further states that an important factor in deciding on sample size is the degree of heterogeneity of the target population. The target population is heterogenous – the individuals speak different L_1's (first language) (see Table 2) – therefore, a

Table 2. Sample group used in the trial run.

Nationality	Examinees		L_1[a]
	No.	%	
Bangladeshi	8	12	Bengali
Bhutanese	2	3	Dzongkha, Nepali
Burmese	6	9	Burmese
Chilean	1	1	Spanish
Chinese	5	7	Chinese
Filipino	9	13	Ilocano, Visayan, Cebuano, Ilonggo, Tagalog
Gambian	1	1	Mandingo
German	1	1	German
Indian	6	9	Punjabi, Marathi, Urdu, Hindi
Indonesian	4	6	Indonesian
Japanese	1	1	Japanese
Korean	1	1	Korean
Mexican	1	1	Spanish
Nepalese	1	1	Nepali
Pakistani	1	1	Urdu
Sri Lankan	8	12	Sinhalese, Tamil, Sinhala
Thai	8	12	Thai
Vietnamese	3	12	Vietnamese
Total	67		

[a] First language or mother tongue.

larger sample would have been desirable. However, only two training programs were going on during the period and many IRRI scholars were too busy to participate.

The first batch of examinees consisted of IRRI scholars working for masteral or doctoral degrees at the University of the Philippines at Los Baños who had been in residence for 6 mo to more than a year. The second batch consisted of cropping systems trainees in residence for about 2 mo, and the third consisted of agricultural engineering trainees who had just arrived at IRRI. The sample represented 18 nationalities (Table 2).

Sixty-five of the 67 examinees had degrees in agriculture or related fields. One had a degree in political science and the other in commerce. Sixty-three had English in high school or college or both, 3 had no English in school but attended an intensive course in English before coming to IRRI, and 1 had no English instruction.

Although the test was planned for 5 h, additional time allowed slow performers to attempt all items. The testing room was spacious, air-conditioned, and well-lighted; the examinees were comfortably seated and had ample working space; the equipment used for the listening comprehension test delivered clear playback of the taped material and the volume was adjusted to the proper level; and examinees were allowed to smoke or take light snacks during the test to ease their tension.

Statistical analysis

1. *Checking and scoring.* Each correct response in the objective tests received a value of 1.

In the cloze test, synonyms were accepted. For example, if the item was "As the disease develops, lesions enlarge and _____ brown," "turn" and "become" were both acceptable.

The tests on following oral directions used weighted scoring. Maxima were 15 points for item 1 and 10 for item 2. Each individual instruction followed correctly was given a point, however, examinee got a .5 deduction for every uncalled-for answer.

In the lecture part, the examinee got 2 points for every correct idea reconstructed. He was not penalized for misspelled words or grammatical errors. A *manufactured* idea, or an idea not explicitly given in the lecture, received no score.

2. *Item analysis.* Item analysis estimates how much information each single item in a test contributes to the information provided by the test as a whole. Item analysis enables the test writer to find out how easy or difficult an item is (index of difficulty), and how well it distinguishes the better students from the poorer ones (index of discrimination).

a. Answer sheets were arranged from highest to lowest and divided into 3 groups – the top 27% (good performers), the bottom 27% (poor performers), and the middle 46%.

b. The correct answers to each item for each group of examinees were tallied and summed, and the sum of the correct answers for all the papers was obtained.

c. The index of difficulty (IDIF) for each item was computed:

$$\text{IDIF} = \frac{\text{TRR}}{\text{N}}$$

where TRR = total right responses, and
N = number of examinees.

d. The index of discrimination (IDIS) for each item was computed by Garett's (6) method. The method considers only the correct responses in the top 27% and in the bottom 27%. First, the number of correct responses in the top 27% and that in the bottom 27% are converted into proportion by dividing those by the number of examinees who attempted the item. The proportions obtained are then located in Flanagan's table (Table 3). The r_{bis} from the intersecting column and row in the table is read. The r_{bis} obtained is the IDIS for the particular item.

e. The numerical values were translated into descriptive values according to Ebel's (4) criteria:

IDIF

.86-1.0	Very easy items
.71-.85	Easy items
.40-.70	Desirable items
.15-.39	Difficult items
.01-.14	Very difficult items

IDIS

.40 and up	Very good items, highly discriminating
.30-.39	Reasonably good but possibly subject to improvement
.20-.29	Marginal items, usually needing improvement

Table 3. Normalized Biserial Coefficient of Correlation as determined from proportions of correct responses in upper and lower 27% of the sample group.[a]

Proportion of correct responses in the lower 27 per cent	Proportion of correct responses in the upper 27 per cent																								
	02	06	10	14	18	22	26	30	34	38	42	46	50	54	58	62	66	70	74	78	82	86	90	94	98
02	00	19	30	37	43	48	51	55	58	61	63	66	68	70	72	73	75	77	79	80	82	84	86	88	94
06		00	11	19	26	31	36	40	44	47	50	53	56	59	61	64	66	68	71	73	76	78	81	84	88
10			00	08	15	21	26	30	34	38	41	45	48	51	54	57	60	63	65	68	71	74	77	81	86
14				00	07	12	18	22	27	31	34	38	42	45	48	51	54	57	60	63	67	70	74	78	84
18					00	06	11	16	20	25	28	32	36	39	43	47	49	53	56	60	63	67	71	76	82
22						00	06	10	15	19	23	27	31	34	38	42	45	49	52	56	60	63	68	73	80
26							00	05	09	14	18	22	26	30	33	37	41	44	48	52	56	60	65	71	79
30								00	04	09	13	17	21	25	29	33	37	40	44	49	53	57	63	68	77
34									00	04	09	13	17	21	25	29	33	37	41	45	49	54	60	66	75
38										00	04	08	13	16	20	25	29	33	37	42	47	51	57	64	73
42											00	04	08	12	16	20	25	29	33	38	43	48	54	61	72
46												00	04	08	12	16	21	25	30	34	39	45	51	59	70
50													00	04	08	13	17	21	26	31	36	42	48	56	68
54														00	04	08	13	17	22	27	32	38	45	55	66
58															00	04	09	13	18	23	28	34	41	50	63
62																00	04	09	14	19	25	31	38	47	61
66																	00	04	09	15	20	27	34	44	58
70																		00	05	10	16	22	30	40	55
74																			00	06	11	18	26	36	51
78																				00	06	12	21	31	48
82																					00	07	15	26	43
86																						00	08	19	37
90																							00	11	30
94																								00	19
98																									00

[a]This table is abridged from J. C. Flanagan's table of normalized biserial coefficients originally for the Cooperative Test Service. This version is taken from Tate (13). Decimal points are omitted. If the proportion of correct responses in the lower 27% exceeds that in the upper, enter the table with the lower 27% proportion at the top and attach a negative sign to the coefficient.

Below .19 Poor items, to be rejected or improved by revision

Of the 376 items, 156 (41%) were desirable, 35 (9%) were difficult, 5 (1%) were very difficult, 125 (33%) were easy, and 55 (15%) were very easy. Two hundred eighty-one (75%) were highly discriminating, 46 (12%) were reasonably good, 19 (5%) were marginal, and only 30 (8%) were poor items.

Retaining and weeding out items. Ebel's (4) criteria for evaluating the IDIF and IDIS, and the discriminating power of the items were used as the principal bases for retaining, revising, or discarding items.

The following items were retained:

- 140 items with an IDIF of .40-.70 and IDIS of .20 or higher,
- 106 items with an IDIF of .71-.85 and IDIS of .20 or higher,
- 18 items with an IDIF of .86-.90 and IDIS of .40 or higher, and
- 27 items with an IDIF of .15-.39 and IDIS of .20 or higher.

Of the 291 items retained, 48% were at a desirable level of difficulty and 85% were highly discriminating. Forty per cent met both of these levels.

Some items which satisfied the criteria on difficulty and discrimination were dropped in favor of another item in the same sentence which tested a more important linguistic component. A part of the listening comprehension test which elicited a true-false response was totally dropped because only 27% of the total number of items were desirable, which could have been due to the susceptibility of the items to chance errors resulting from guessing.

A summary and analysis of retained and discarded items are given in Tables 4 and 5.

Assembling the final form. The retained items in each part, except in the maze and cloze tests, were arranged from the easiest to the most difficult, based on their difficulty indices.

Retained items with a marginal discriminating power (.20-.29) were revised, either in the stem (the initial part of an item, either a partial sentence to be completed, a question, or several statements leading to a question or an incomplete phrase) or in the options (alternate-response choices intended to attract students who do not know the right answer), and in some cases, in both. In revising an item, the writer noted the part that distracted even the good pe formers toward the wrong choices.

The following example illustrates how an item with a marginal discrimination index was revised:

Item in the trial run:

Precipitation (rainfall plus snow) is the primary source of water in agriculture. Which of the following functions does the phrase "rainfall plus snow" fulfill?

a. identifying
b. describing
c. defining

Table 4. Summary of items retained and discarded in each difficulty level.

Difficulty level	Number of items				Total no. of items	Total			
	Vocabulary and Reading		Listening Comprehension			Retained		Discarded	
	Retained	Discarded	Retained	Discarded		No.	%	No.	%
.86-1.00 (Very easy items)	15	25	3	12	55	18	33	37	67
.71-.85 (Easy items)	85	5	22	13	125	107	86	18	14
.40-.70 (Desirable items)	110	6	29	11	156	139	89	17	11
.15-.39 (Difficult items)	26	6	1	2	55	27	77	8	29
.01-14 (Very difficult items)	0	5	0	0	5	0	0	5	100
Total	236	47	55	38	376	291	77	85	23

Table 5. Analysis of retained items.

Difficulty Index			Discrimination Index		
Criteria	Total no. of items	%	Criteria	Total no. of items	%
.86-1.00 (very easy items)	18	6	.40-up (very good items)	248	85
.71-.85 (easy items	107	37	.30-.39 (reasonably good items)	32	11
.40-.70 (desirable items)	139	48	.20-.29 (marginal items)	11	4
.15-.39 (difficult items)	27	9	Below .19 (poor items)	0	0
.1-.14 (very difficult items)	0	0	Zero to negative items	0	0
Total	291	100	Total	291	100

The writer noted that the stem was not clear to most examinees. The stem was, therefore revised, to read: *Precipitation (rainfall plus snow) is the primary source of water in agriculture.* The sentence defines precipitation as _____.

a. the primary source of agriculture
b. water supplied to irrigated rice
c. rainfall and snow

Sample items and instructions of the final form are given in the appendix.

Computing the reliability coefficient. Pilliner (11) states that it is not enough that a test should discriminate. It should discriminate consistently. Scores on it must be reliable. This means that when the procedure is repeated, the scores should not be subject to fluctuations, and hence, the persons obtaining these scores should maintain their standings relative to one another.

The reliability of a test depends on how sharply the items discriminate between good and poor students, how many items there are, how similar the items are with respect to the ability measured, and how much the students differ from one another in the ability being measured.

Ebel (4) enumerates three factors contributing to the reliability or unreliability of a set of test scores:

- the appropriateness and definiteness of the task,
- the consistency or stability of a student's ability to perform the tasks presented in the test, and
- the consistency and objectivity of the person scoring the test.

A reliability coefficient estimates the correlation between one set of scores on a particular test for a particular group of examinees and an independent set of scores on an equivalent test for the same examinees. The higher this coefficient, the more consistently the test is measuring whatever it does measure.

The reliability coefficient of the final form of the test was computed using the Kuder-Richardson Formula 20:

$$r = \frac{k}{k-1} \left[1 - \frac{\Sigma pq}{\sigma^2} \right]$$

where:
k = number of items in the test,
Σ = the sum of,
p = the proportion of correct responses to one item,
q = the proportion of incorrect responses to one item, and
σ^2 = the variance of the scores on the test.

The Kuder-Richardson Formula 20 measures the internal consistency, or homogeneity, or scalability, of the test material. If the items on a test have high intercorrelations with each other and are measures of much the same attributes, the reliability coefficient will be high. If the intercorrelations are low, either because the items measure different

attributes or because of the presence of error, then the reliability coefficient will be low.

The reliability coefficient of the test was 0.99045, indicating the test items are highly intercorrelated.

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

The test is an evaluation instrument in English for Special Purposes. It measures the reading and listening comprehension in English of non-English applicants for training at IRRI. The test focuses on receptive skills, because an IRRI trainee spends about 75% of his time listening to lectures or seminars and reading handouts, notes, and other reading materials. Writing and speaking skills are not considered as critical as one's ability to comprehend written and spoken materials.

The test can be considered satisfactory because the items are moderately difficult (48% have a difficulty index of .40-.70) and highly discriminating (85% have a very good discrimination index). A test is considered ideal when more than 25% of the items discriminate well and 25% are moderately difficult (The University Examination Service, University of Iowa, as cited in 8). The test is valid and reliable because the items have been constructed around the receptive skills required in the examinees' future role as agricultural trainees and the test's reliability coefficient is 0.99045. Expertly constructed educational achievement tests often yield reliability coefficients of .90 or higher (4).

Most of the desirable items in the reading test were in ordering sentences (50%) and in the cloze test (49%). Most of the desirable items in the listening comprehension test were in question comprehension. Most of the highly discriminating items in the reading test were in the reading comprehension (82%) and cloze (89%) tests. In the listening comprehension test, most of the highly discriminating items were in following oral directions, lecture, and question comprehension.

The cloze test seemed to provide a reliable measure of the global reading comprehension skills of the examinees because it calls for the appropriate linguistic knowledge, textual knowledge, and knowledge of the world. Only the good performers did well in the cloze test. The examinees had the most difficulty with completions involving devices of contrast, time expressions, defining and nondefining relative clauses, and grammatical agreement. The easiest were those involving enumerations, comparisons, additions, and generalizations.

The following steps can be taken to improve the test:

1. The test should be administered to a larger sample, composed preferably of trainees who have just arrived at IRRI.
2. The test should be further reviewed for possible revisions and improvement. The test should then be retested and statistically analyzed.
3. The test can be given to a group of nonagriculturists, to determine to what extent it measures knowledge of agriculture.
4. Other methods of estimating the reliability coefficient of the test should be used.
5. The test should undergo a long period of testing-retesting for standardization before it is used as a screening tool for applicants for training at IRRI.
6. The concurrent and predictive validity of the test should be established. (Concurrent validity is concerned with the relation of test scores to an accepted contemporary criterion of performance on the variable which the test is intended to measure. Predictive validity is concerned with the relation of test scores to measures on a criterion based on performance at some later time.) Does the test accurately measure what is intended to measure? Can an individual's score in the test be used to predict his performance later as an IRRI trainee?

After the test papers had been checked and scored, a list of the good performers in the test was sent to the trainers in their respective training programs to find out trainee performance. The trainers confirmed that the good test performers were also the good performers in the training course.

A more reliable indication of the predictive validity of the test was obtained with the second batch of examinees who completed their cropping systems training 2 mo after they took the English proficiency exam. The official list of the graduating group was checked against the ranking in the English exam. The 23 who completed the course with distinction were either good performers or in the middle 46% in the English test. The poor performers in the course were also the poor performers in the English test. This seems to indicate a correlation between one's score in the English test and his performance as an IRRI trainee. However, this is not conclusive. Further studies should be conducted to determine correlation between examinees' scores in the English test and performances as IRRI trainees.

REFERENCES CITED

1. Alderson, J. C. 1981. Report of the discussion on testing English for specific purposes. *In* Issues in language testing. British Council.
2. Carroll, B. 1982. Testing communicative performance. (Course handout)
3. Cuevas, S. L. 1971. A proposed English proficiency test in pronunciation for six Asian student groups. Special problem, University of the Philippines, Diliman, Quezon City.
4. Ebel, R. L. 1965. Measuring educational achievement. Prentice-Hall, Inc., New Jersey.
5. Francisco, V. V. 1972. A proposed English proficiency test in vocabulary for six selected Asian student groups. Special problem, University of the Philippines, Diliman, Quezon City.
6. Garett, H. E. 1958. Constructing evaluating instruments. Longmans, Green and Co., New York.
7. Ingram, E. 1968. Item analysis. Pages 192-203 *in* Language testing symposium. A psycholinguistic approach. Alan Davies, ed. Oxford University Press, London.

8. Mackay, R., and A. J. Mountford. n. d. The teaching of English for special purposes: theory and practice. *In* English for special purposes. (Course handout)
9. Manilay, F. A. 1971. A proposed English proficiency test in grammar for six Asian student groups. Special problem, University of the Philippines, Diliman, Quezon City.
10. Nartea, R. V. 1980. Development and standardization of an English proficiency test for fourth year high school students of agricultural schools. Dissertation, University of the Philippines, Diliman, Quezon City.
11. Pilliner, A. E. G. 1968. Subjective and objective testing. Pages 19-35 *in* Language testing symposium. A psycholinguistic approach. Alan Davies, ed. Oxford University Press, London.
12. Strevens, P. 1980. English for special purposes: an analysis and survey. Pages 458-472 *in* Readings in English as a second language for teachers and teacher trainees. Little, Brown and Co., Boston.
13. Sunkatup, N. 1981. The development of an integrative pragmatic English proficiency test for fourth year student teachers in Thailand. Special problem, University of the Philippines, Diliman, Quezon City.
14. Tate, M. W. 1955. Statistics in education. The MacMillan Co., New York.
15. Villar, A. 1971. The development of an integrative proficiency test in English for fourth year secondary school students of Nueva Vizcaya. Special problem, University of the Philippines, Diliman, Quezon City.
16. Walsh, V. 1982. Reading scientific texts in English. English Teaching Forum, July 1982, p. 24-37.

Appendix. Sample items from the final form.

READING TEST

I. Vocabulary. Choose the word that means nearly the same as the underlined word in the sentence. Write only the letter of your choice.

1. *Plug* the tube mouth with cotton.
   a. clean b. dry c. stuff d. provide

II. Ordering Sentences. The sentences of the following paragraph are not in the correct order. Rearrange the sentences so that they become well-ordered paragraph. Choose the correct sentence order from the options given. Write only the letter of your choice.

1. The ox-drawn wooden plow, the sickle, the flail, and threshing by animal hooves are found on almost every farm.
2. Yet many of the tools and practices developed in those ancient times still predominate.
3. Modern farm mechanization has essentially bypassed Egyptian agriculture.
4. There are many reasons for this, but the greatest constraint perhaps is the lack of local fabricators.
5. Egypt has a 5000-year history of agricultural cultivation.

The correct sentence order is:
a. 1-2-4-3-5 b. 3-2-1-4-5 c. 5-4-1-2-3 d. 5-2-1-3-4

III. Maze. Some words have been deleted in this passage. Read the passage and choose from the alternatives the word that makes sense according to the meaning of the passage.

IRRI studies throughout Asia indicate that reliability, durability, and low cost are the primary concerns of potential buyers of farm machineries. In response, IRRI developed two machines that [1. a. offer / b. include / c. specify / d. fit] a wide variety of conditions and [2. a. increase / b. reduce / c. require / d. restrict] labor and power constraints, which often [3. a. reduce / b. increase / c. allow / d. cover] the turnaround time between crops.

IV. Sentence Comprehension. Choose the word or phrase that best completes the sentence or that best relates the sentences in a number. Choose only one answer and write only the letter of your choice.

1. Granulated clayey soils not only are well-aerated and well-drained, they _____ can retain sufficient water for plant use.
   a. therefore b. also c. similarly d. however

V. Reading Comprehension. This is a test of how well you read and understand passages. Read the following passage and then answer the questions after each passage. Write only the letter of your choice.

[1]In cooperative studies with the International Centre of Insect Physiology and Ecology, the extract of an indigenous plant was found effective against numerous insect pests. [2]Even extremely small quantities of the extract repelled insect pests, reduced their feeding, disrupted embryonic and larval growth, and decreased survival and oviposition. [3]Sprays of the plant extract reduced BPH nymphs and adult populations in the field.

1. Which of the following statements is true according to the passage? _____
   a. A small amount of the extract was ineffective against insects.
   b. The extract had an offensive smell.
   c. Insects fed less on plants sprayed with the extract.
   d. Insects that fed on sprayed plants became sterile thereby reducing their population.
2. *Their* in sentence 2 refers to _____.
   a. plant extracts
   b. quantities
   c. nymphs
   d. insect pests

VI. Cloze. Some words have been deleted in the following passage. Read each passage carefully and insert whatever word makes sense according to the meaning of the passage. The word should be grammatically correct. Use only one word for each space. Read the passage at least once before you start to write.

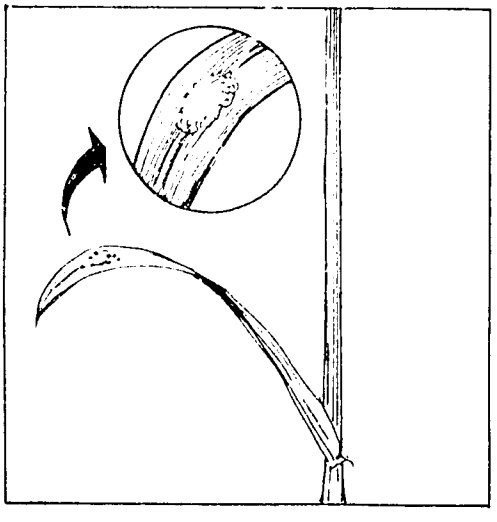
We compared the abundance and survival of brown planthopper and two important predator species in two IRRI fields – one was treated with insecticide and one nearby was left untreated. The brown planthopper population in the untreated field remained small and survival was very low as expected, _____(1)_____ in the treated field pest, density _____(2)_____ survival increased, ultimately causing hopperburn. _____(3)_____ of insecticide treatment, the abundance _____(4)_____ natural enemies such as predatory spiders and ripple bugs decreased.

LISTENING COMPREHENSION

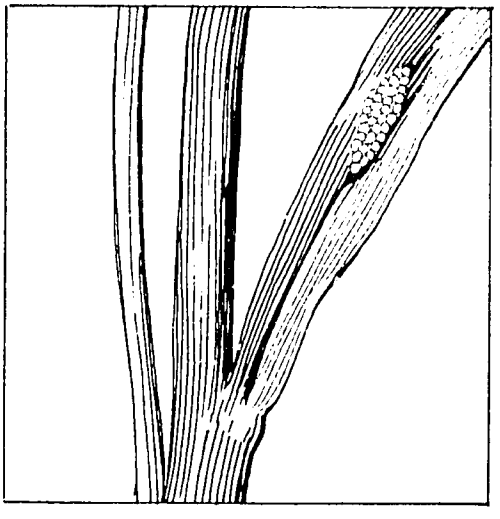
I. There are _____ problems in this test. Listen carefully as the examiner speaks each problem and study the three pictures that go with each problem. On the basis of what the examiner says, choose the picture that you think is correct. Write the letter of your choice on the separate Answer Sheet.

Problem (TAPED): The eggs are laid near the tip of the leaf blade.

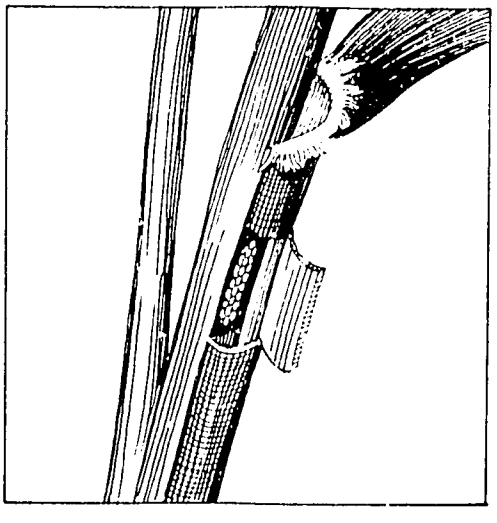
Illustrations on the Exam Booklet:



A



B



C

II. Question Comprehension. You will hear _____ questions. Listen carefully as the examiner reads each. Look at the four alternatives written on your exam booklet. Choose the one that you think answers the question correctly. Write the letter of your choice on the separate Answer Sheet.

Question (TAPED): Why are the nymphs and adults of rice bugs difficult to see in rice fields?

Options (printed on Exam booklet):

a. Rice bugs can be detected because they emit the typical rice bug odor.
b. Their color resembles that of the rice plants.
c. They cause great damage to rice fields.
d. Their population usually increases at the end of the rainy season.

III. Following Spoken Directions. This is a test of how well you can understand and how accurately you can follow spoken instructions. You will hear a series of instructions. Do exactly what you are asked to do. You will have 5-10 seconds to follow each instruction, depending on the difficulty of the instruction. The blank Answer Sheet is for your answers.

Variety	Nodal tillers (no.)[a]		
	M	P	S
Laki 550	1.0	5.0	6.5
Lalamon 245	0.8	5.3	3.3
Gowai 476	0	2.0	0
Shuli	0	0	1.0
Nam Sagui	0	0	0

[a]M = , P = , S =

TAPED INSTRUCTIONS:
Study the table.
1. Extend the line below *Nodal tillers* and add a fourth column. Label that column *Total*.
2. Add the numbers under the columns M, P, and S for each variety and write the total under the column *Total*.
3. Complete the footnote by indicating what M, P, and S mean. M stands for main, P for primary, and S for secondary.

IV. You are going to hear a lecture on _____. You are being tested on your ability to listen and understand. After the lecture you will be asked _____ questions about it. Look at the Notes Sheet. The Notes Sheet is for any notes you may wish to take while you are listening to the lecture. The points listed in the Notes Sheet can aid you in determining what notes to take. If you wish to take notes other than those listed, you may do so. Your notes will not be marked or graded. The questions that you will be asked after the lecture will be on the points listed in the Notes Sheet.

You will also be asked to summarize the lecture by reconstructing the notes you have taken during the lecture. You will be graded on the basis of how much correct information from the lecture you have taken. Each correct point is worth 2 points. Total possible score for your summary is 20.

The International Rice Research Institute
P.O. Box 933, Manila, Philippines

Stamp

ISSN 0115-3862

Other papers in this series

TITLES OF NUMBERS 1-45 ARE LISTED ON THE LAST PAGE OF NO 46, THOSE OF NUMBERS 46-70 ARE ON THE LAST PAGE OF NO 71-80

No. 71 The development and diffusion of rice varieties in Indonesia

No. 72 Levels of resistance of rice varieties to biotypes of the BPH, *Nilaparvata lugens*, in South and Southeast Asia

No. 73 Growing season analyses for rainfed wetland fields

No. 74 San Bartolome: beyond the green revolution

No. 75 Pathotypes of *Xanthomonas campestris* pv. *oryzae* in Asia

No. 76 Focusing field research on future constraints to rice production

No. 77 An international survey of methods used for evaluation of cooking and eating quality of milled rice

No. 78 Research on algae, blue-green algae, and phototrophic nitrogen fixation at the International Rice Research Institute (1961-81), summarization, problems, and prospects

No. 79 Seed-derived callus culture for selecting salt-tolerant rices

No 80 Economic limitations to increasing shallow rainfed rice productivity in Bicol, Philippines

No. 81 Irrigation system management research and selected methodological issues

No. 82 Interdisciplinary challenges and opportunities in international agricultural research

No. 83 Comparative analysis of cropping systems: an exploratory study of 14 rainfed sites in the Philippines

No. 84 Rapid generation advance of rice at IRRI

No. 85 Physicochemical characterization of iron-toxic soils in some Asian countries

No. 86 New rice technology, intrarural migration, and institutional innovation in the Philippines

No. 87 RICEMOD: a physiologically based rice growth and yield model

No. 88 Sensitivity tests of the environmental variables in RICEMOD

No. 89 Sensitivity tests of the crop variables in RICEMOD

No. 90 New rice technology and labor absorption: comparative histories of two Philippine rice villages

No. 91 Calculating the private benefits of farm machinery: a microcomputer application

No. 92 Cropping systems research in the Pangasinan Project

No. 93 Estimating risk of fertilizer use in rainfed rice production

No. 94 Sensitivity tests of the environmental variables in IRRIMOD

No. 95 Sensitivity tests of the crop and management variable in RICEMOD

No. 96 Fertilizer transfer to floodwater during deep placement

No. 97 Interaction between fertilizer and weed control methods in Philippine upland rice: estimates from farmer's fields

No. 98 Training needs of information services in agricultural research and educational organizations in Asia: a 9-country survey

No. 99 Soil sickness caused by continuous cropping of upland rice, mungbean, and other crops

No. 100 Changes in input use and grain yields in lowland rice farms in three Philippine provinces

No. 101 The economics of hybrid rice production in China

No. 102 Rice ratooning

No. 103 Growth and development of rice grown in deep water

No. 104 Faridpur: a computer-assisted instruction model for rainfed lowland rice