DOCUMENT RESUME

ED 462 861 FL 027 201

AUTHOR Bolton, Sandra; Johnson, Diane; Lyons, Caryl; Gaies, Stephen J.

TITLE Interrelationships among Skill Areas in Standardized Assessment.

PUB DATE 2001-03-01

NOTE 33p.; Paper presented at the Annual Meeting of Teachers of English to Speakers of Other Languages (35th, St. Louis, MO, February 27-March 3, 2001).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Communication Skills; Elementary Secondary Education; *English (Second Language); Foreign Countries; *Language Proficiency; Language Skills; Listening Skills; Reading Skills; Second Language Learning; *Standardized Tests; Writing Skills

IDENTIFIERS *Japan

ABSTRACT

This paper describes the development of the Benesse Proficiency Test of English Communication, which was created for the Benesse Corporation of Japan. The Benesse test was developed in response to Japan's movement toward more communicative English language teaching and testing. It was intended to test language used in authentic communication. The paper also discusses statistical data from the first operational test administration. The data examined are from 35,000 Japanese students who took the advanced level operational form 1 in 1998. Data were analyzed to evaluate the performance of the instrument, focusing on the observed relationships, or intercorrelations, among the skills of listening, reading, and writing. Results indicated that the intercorrelations among the subtests were similar to those of other tests, which provides validation of this new test. (Contains 18 references and 3 tables.) (SM)

Interrelationships among Skill Areas in Standardized Assessment

Sandra Bolton
Diane Johnson
Caryl Lyons
Stephen J. Gaies

Paper presented at the Annual Meeting of Teachers of English to Speakers of Other Languages (35th, St. Louis, MO, February 27-March 3, 2001).

FL

INTERRELATIONSHIPS AMONG SKILL AREAS IN STANDARDIZED ASSESSMENT

TESOL 2001
St. Louis, Missouri
Thursday, March 1, 2001

Sandra Bolton, ACT, Inc., Iowa City, IA
Diane Johnson, ACT, Inc. Iowa City, IA
Caryl Lyons, ACT, Inc., Iowa City, IA
Stephen J. Gaies, University of Northern Iowa

I. Introduction

This paper describes the Benesse Proficiency Test of English Communication, which ACT, Inc. developed for the Benesse Corporation of Japan, and also discusses the statistical data from the first operational test administration. The data examined are from approximately 35,000 Japanese students who took the Advanced Level Operational Form 1 in 1998. At that time, a Basic Level Test was also administered to a similar number of students and yielded similar data. For this paper, we have chosen to limit our study to data from the Advanced Test.

Why the Benesse test was developed

In 1989 and 1990, the Japanese Ministry of Education, *Mombusho*, released a new curriculum of English language learning that emphasizes communicative language teaching. Traditionally, English language education in Japan has emphasized reading and grammar, in part, perhaps, because of the ease of standardized assessment, and university entrance exams have predominantly focused on reading and writing tests.

Reading has usually been taught in Japan through translating, memorizing vocabulary, and sometimes reading aloud repetitively. Japanese teachers of English have usually explicated reading texts in detail, primarily using Japanese rather than English. Writing has most often consisted of single sentences translated from an original Japanese version (LoCastro, 1996). None of these strategies fit into a model of communicative language teaching.

Several factors motivated the Ministry of Education to create the new curriculum. As more Japanese students studied abroad and went on homestays, it was recognized that the English taught in school was not practical for common communication, that is, listening and speaking. It became evident that students need to be able to understand spoken English, and to speak English well themselves, if they are to succeed in business, government, science, and many other fields, as well as to succeed in study abroad in English-speaking countries. Furthermore, Japanese students' test scores on the TOEFL and TOEIC were consistently lower than those of students in other Asian countries (Mulvey, 1999). In response, the Ministry of Education began changing its guidelines for English proficiency. And now some Japanese universities are altering their entrance requirements to reflect the acknowledged need for communicative English, and have begun to test speaking and listening skills. For example, the prestigious University of Tokyo recently introduced a listening comprehension component in its entrance exam (LoCastro, 1996).

Consequently, Japanese secondary schools are under pressure to provide more instruction in listening and speaking, and in writing for communication. Teachers are being required to teach more communicatively, but many have never been taught communicative methods of teaching. And some teachers are unsure of their own communicative English skills.

Communicative nature of the Benesse test

Let's examine what's meant by communicative testing and look at the communicative nature of the Benesse test. Among the definitions of communicative language teaching and testing are those mentioned by Kitao and Kitao (1996): (1) students should develop the ability to use language in real-life situations; (2) tests should reflect communicative situations in which testees are likely to find themselves or social situations in which they might be in a position to use English; and (3) the receptive skills of listening and reading should emphasize understanding the communicative intent of the speaker or writer. The Australian Board of Senior School Secondary Studies notes that communicative testing should use authentic texts and give students the opportunity to speak and write from their own experience (Sato and Kleinsasser, 1999).

The Benesse test was developed in response to the new communicative curriculum. It aims to test language that is used in authentic communication, and to test this language in a more communicative manner. It is, of course, not always possible to make standardized language assessments entirely communicative when multiple-choice questions are used. But it is possible to enhance the communicative nature of a test through the use of

authentic materials and tasks, which students might encounter in the real English-speaking world. The three subtests contain various authentic elements.

The listening portion of the test reflects authentic American speech patterns, syntactically and phonetically, and in a large percentage of the items, replicates real-life situations. The reading items are based on reading skills used in real life—skimming, scanning, and deciphering the meaning of words from context. And the majority of the reading texts are based on authentic materials originally written for native English speakers. In general, the Benesse test focuses on the meaning intended by the language rather than on the structure of the language.

A direct writing assessment is the third Benesse subtest. A direct writing assessment necessarily involves communication between the writer and the rater, or scorer. However, the Benesse writing assessment is communicative primarily because it evaluates content rather than the grammatical accuracy of student writing.

II. Subtests of the Benesse test

All items on the Benesse test are pretested, and, on the basis of the pretest results, items are selected for constructing the operational test forms. The Listening and Reading Tests have nearly the same number of items, and receive equal weight in the composite score. Both Listening and Reading scores are scaled to a 32-point scale, which is not so coarse that score precision information is lost (too many raw score points mapping to the same scale score), or so fine that adjacent score points reflect trivial differences in ability

relative to measurement error. The Writing Test consists of two essays that are added together to make a 12- point scale with ½ point intervals, which is then rescaled to a 16-point scale and included in the composite score. Students also receive individual subtest scores. The various subtests of the Benesse test are described below to provide a context for understanding the statistics that follow.

Below are examples from each part of the Benesse test.

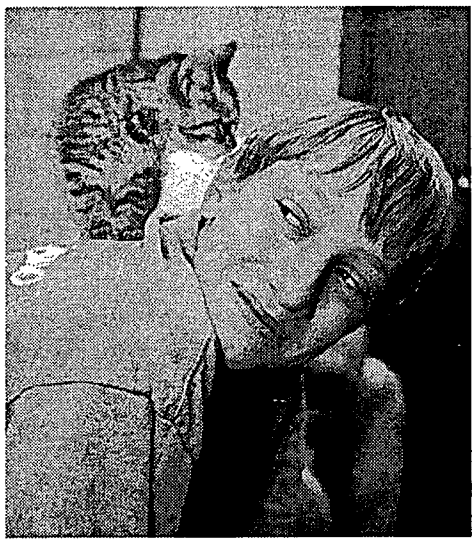
The Listening Test

The Listening Test, which is made up of four sections—Parts A through D—is 25 minutes long. There are 40 multiple-choice listening items (10 items per part), enough items to gain an accurate measure of students' listening abilities. Moreover, careful adherence to an approved vocabulary list, assembled from words and idiomatic phrases found in English language textbooks used in secondary schools in Japan, helps ensure that the test will not be too difficult. This is important since too many low scores do not permit adequate differentiation among lower-level students. There is no repetition of spoken stimuli—conversations and monologues—or the response options that students hear on tape, but students may take notes in the margins of the test book.

The Listening Test tries to simulate authentic spoken American English, so in the stimuli and options that are recorded, words are naturally linked, and some contractions and phonetic changes (assimilations and reductions) are used. Graphics play a significant role in the Benesse Listening Test—half the test items include either photos or

illustrations—in an effort to come closer to real-life listening situations, which are usually accompanied by visual cues for the listener.

Listening Part A requires students to choose the best statement out of three recorded statements they hear to match with each photograph. The students see:



The students hear:

A. The cat is jumping into the arms of the boy.

B. The cat is lying on its back in the sunlight.

C. The cat is sniffing the child's ear.

The correct answer is [C], "The cat is sniffing the child's ear."

With this item type, the photographs permit descriptive language to be tested along with some inferencing skills.

In **Listening Part B**, students hear a short question and three responses, and then quickly choose the correct response to the question. Here is an example:

Did you have fun at the party?

A. I was funny.

B. It was great.

C. There was no phone.

The correct answer is [B]; "It was great." The two-turn, social-interactional conversations in Part B use ritualized, idiomatic turns, and high-frequency conversation topics.

For **Listening Part C,** students "perform" real-world tasks where English would naturally be used between a Japanese student and a native English speaker. The situation in which the discourse takes place, the scenario, is described with a few sentences printed in Japanese in the testbook and also heard on the recording. Students can read the pair of questions associated with the scenario before they hear the stimulus.

The scenario is printed here in English:

A group of American exchange students living in your city is having a party. Your friend, an American boy in your class, invited you, too. Your friend, who is busy getting food ready, asks you to take some things to people at the party. Listen carefully so you can find the right person and take the right thing.

Students listen to the stimulus while they look at an illustration with four options identified. The students hear:

[American boy (M), Japanese girl (F)]

M: My brother asked for some crackers. Would you take him some?

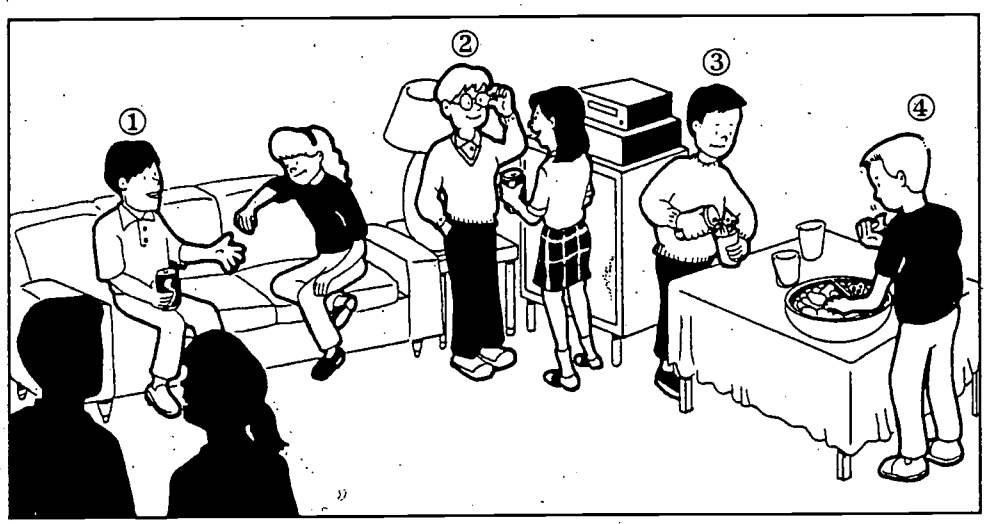
F: Sure. Which one is he?

M: He's the boy in the sweater.

F: Is he pouring something into a glass?

M: Yes. Also, Nancy isn't feeling well and wants a glass of water. Could you take her some?

F: OK. I know Nancy. I met her yesterday at the coffee shop.



Which one is your friend's brother?

A. 1

B. 2

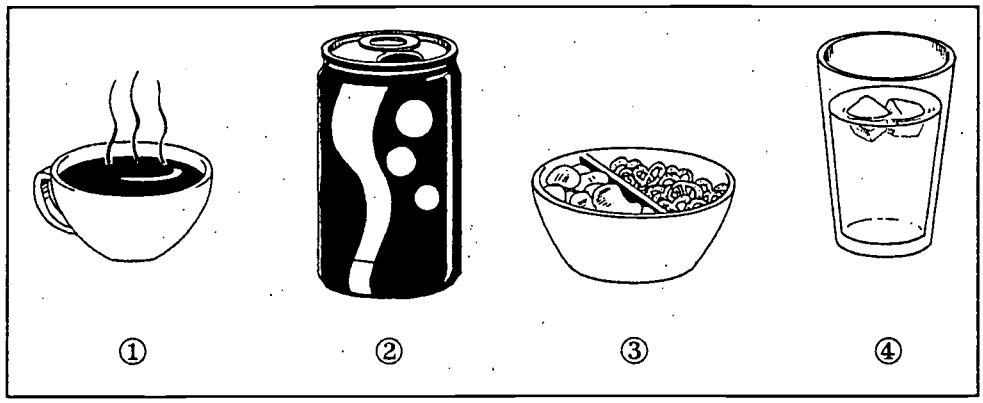
C. 3

D. 4

Then students choose the correct answer to the question printed in the testbook.

Question #1: Which one is your friend's brother? The correct answer is [C], #3, the boy in a sweater who is pouring a drink.

There is a second graphic for Question #2:



What will you take to Nancy?

A. 1

B. 2

C. 3

D. 4

The correct answer is [D], #4, a glass of water.

Most of the illustrations depict a human figure that represents the student who will carry out the task. In addition, young speakers are used in the recordings to represent the Japanese student (the test taker) who must perform the task. This helps draw students taking the test into the dialogues by making them more relevant. To answer the questions, students must recognize discrete information, or details, as well as be able to understand the interaction among the details.

Listening Part D asks students to find specific information in short conversations (four- and six-turn dialogues) and monologues (a single-speaker giving information). The questions for these longer stimuli require comprehension of both main ideas and details. This is an example of a Part D item where students read the question before hearing the stimulus.

M: Who was the woman you were with yesterday, Becky?

F: My mother. We were shopping for a birthday present for my father.

M: You two really look alike.

F: Do you think so?

"What does the man say about Becky and her mother?"

A. That they like shopping

B. That they like her father

C. That they bought a birthday present

D. That they look similar.

The correct answer is [D], "That they look similar."

The Reading Test

The Reading subtest is basically a traditional test of reading; however, the materials are nearly all adaptations of authentic materials. The Reading Test is communicative in that it uses these authentic materials and that it asks students to use real-life reading skills such as skimming and scanning in addition to the usual close-reading skills. Also, the Reading Test uses basically the same vocabulary list as the Listening Test.

The Reading Test has three parts. **Reading Part A** tests vocabulary in a one- or two-sentence context. The items are specifically constructed to test vocabulary knowledge rather than grammatical knowledge.

> Although everybody said that the baseball player was ill, he was _____ on vacation. Only his coach knew about his vacation plans.
>
> A. generally
>
> B. actually
>
> C. hardly
>
> D. fully

The correct answer is [B], "actually." As you can see, the distractors are usually the same part of speech, in this case adverbs, so that the item is actually testing knowledge of a vocabulary word within a brief context.

Reading Part B has two types of items. The first type consists of short passages (75 to 100 words) that students read quickly to determine the main idea.

Part of an underwater volcano close to Hawaii fell down during the summer, according to ocean scientists. Scientists could see how islands are made on the bottom of the ocean. This is the first time they have been able to watch an island forming. This new island is located about 27 kilometers (km) off the coast of the island of Hawaii. It sits on the bottom of the ocean, 5,400 meters (m) below the surface and rises to 900 meters below the surface. Scientists think the island won't show above the water for another 50,000 years.

What is the main idea of this passage?

A. A new island is forming near Hawaii

B. The ocean is 5,400 m deep near Hawaii.

C. Scientists are studying the coast of Hawaii.

D. A volcano is erupting near Hawaii.

In this case, the paragraph is about scientists finding a new island that is forming near Hawaii. So the correct answer is [A]. Distractors may be facts from the article that are true but are not the main idea, such as [B], "The ocean is 5,400 meters deep near Hawaii."

The second type of Reading Part B item is based on authentic materials, such as schedules, advertisements, and brochures, which students scan to locate specific information. In this case, the authentic material is a schedule for riverboat rides. Students are directed to read the two questions first and then scan the stimulus material for the answers, rather than to read the entire stimulus. In most cases, students need to access at least two pieces of information to answer a question. (Stimulus on next page.)

RIVER CITY RIVERBOAT SCHEDULE

DINNER ENTERTAINMENT: Features a buffet of our award-winning prime rib and savory baked chicken. Also live entertainment and a 2½-hour cruise on the Missouri River.

Saturday 6:00 to 8:30 p.m. . . . . Adult $29.95

Tuesday, Wednesday, Thursday
7:30 to 10:00 p.m. . . . . Adult $26.95

Sunday 5:00 to 7:30 p.m. . . . . Adult $26.95

FRIDAY FAMILY FUN NIGHT: A cruise for the whole family with live entertainment and a buffet dinner.
7:00 to 9:00 p.m. . . . . Adult $19.95
Child $10.00

GOSPEL ENTERTAINMENT: The whole family will enjoy our 2-hour live Gospel music cruise featuring our chef's famous southern fried chicken buffet.
Monday night only, May through October
7:00 to 9:00 p.m. . . . . Adult $19.95

BRUNCH: A delicious brunch buffet, live entertainment, and wonderful 2-hour cruise on the Missouri River.
Saturday and Sunday only
11:30 a.m. to 1:30 p.m. . . . . Adult $19.95

MOONLIGHT ENTERTAINMENT: Start your weekend right on a romantic cruise under the stars, featuring live entertainment Saturday nights. April through October
9:00 to 11:00 p.m. . . . . Adult $10.00

SIGHTSEEING: Enjoy our 1-hour sightseeing cruise aboard the River City Riverboat. Our captain's narration includes river tales and truths of yesteryear as well as today. Concession stand available.
DAILY June through August
Saturday and Sunday: March through May and
September through December
2:00 to 3:00 p.m. . . . . Adult $7.50
Child $3.75

Prices and times subject to change without notice.
All applicable taxes apply.
Kids 3 and under free with paid adult.

About how much would it cost a husband and wife and their 8-year-old daughter to enjoy Family Fun Night on the River City Riverboat?

A. About $25.00

B. About $40.00

C. About $50.00

D. About $60.00

Looking at the stimulus again, you can see that students must first scan to locate "Family Fun Night," and then find the cost for adults and for children. Two adults at about $20 each and 1 child at about $10 means that the approximate cost for all three would be $50, so [C], "About $50," is the correct answer.

What is Gospel Entertainment?

A. A sightseeing tour

B. A prime rib dinner

C. Romance under the stars

D. A type of live music

Students must scan to find "Gospel Entertainment" and then look carefully at the description to see that it refers to two hours of "live music." So the correct answer is [D], "A type of live music."

Reading Part C has three 300- to 325-word passages, each accompanied by 5 questions. These are mostly traditional types of reading questions, such as recognizing main ideas, understanding details and relationships, and figuring out vocabulary from context. The passages are adapted from authentic published materials such as fiction and nonfiction books, short stories, and newspaper and magazine articles. Each test form includes three passages, one on each of the general topics of humanities, fiction, and science.

Here is one example of a Part C passage and two of the items that accompany it. This passage is a fiction passage adapted from the novel *Face to Face,* by Marion Dale Bauer, about a group of people going down the Arkansas River in a boat.

Here is the adapted passage:

Michael stood on the bank of the Arkansas River, listening to the safety talk his father was delivering. When he looked over at the river, it was hard to see why they needed the instructions. The Arkansas River looked shallow and narrow, not any bigger or faster than the stream near the farm at home.

His father and Carmen stepped into the boat and Michael pushed the boat into the river. The water was cold, a bone-chilling cold that he hadn't expected. The fast-moving water seized the boat immediately, although his father was directing their movement with firm strokes of his paddle.

They hit a series of small waves in the water and the boat seemed to move with the changing surface of the river. Michael reached for the safety line that ran across the bow of the boat. When he looked back at the boats following theirs, the other people seemed to be having a good time. No one was hanging on to the safety lines. Except for the people paddling, everyone was really relaxed, laughing and calling to one another.

Suddenly the river seemed to bend, narrow, and divide over a huge rock, all in the same spot. The water roared and splashed into the boat, stinging cold. Pulling his paddle against the water, Michael was held firmly in the boat, until it brushed against a rock and he fell over into the soft bottom of the boat. He straightened himself up when the river became quiet again and they floated on silently, except for the sound of paddles in the water.

Michael heard the next rapids long before he could see them. The water rushing over the rocks sounded like a train. The boat slipped through the roaring, foaming water, past huge rocks waiting to catch them. They came out on the other side as smoothly and easily as if they had just gone down a children's slide.

What did Michael notice about people in the other boat?

A. Everyone was happy and relaxed.

B. Everyone was working hard.

C. Everyone seemed to be afraid.

D. Everyone seemed to be very cold.

Students need to locate the information about other people. The correct answer is [A]. The word "relaxed" is used in the passage. Students must make the inference that if people are "laughing and calling to one another," they are happy.

Based on this passage, what does the word "paddle" mean?

A. The uneven surface of the water

B. A tool used to move the boat

C. To hang on to a rope

D. To sit up straight

Students locate the phrase "directing their movement with firm strokes of his paddle" to determine that the correct answer is [B].

The Writing Test

Students write two short essays on this test that involve two types of writing: descriptive and persuasive. The total writing time is 25 minutes. For Part A, the descriptive writing, students are asked to compare two photographs, noting the similarities and differences between them. For Part B, the persuasive writing, students agree or disagree with a statement, giving as many reasons as they can for their opinion.

Each essay is scored by two [native] speakers of American English using a six-point modified holistic rubric. Scores from the two essays are added together to make a 12 point scale. The scoring features in the rubric cover traits common to most writing assessments such as fluency or development, organization, vocabulary, and language elements such as sentence structure and word choice. Mechanical errors are considered only to the extent that they may interfere with meaning or understanding.

As was mentioned earlier, this writing test differs from other writing assessments in the way that raters are trained to look for development of ideas rather than accuracy of expression. Raters are explicitly taught to ignore "local" errors, that is, ones that do not interfere with the communication of ideas, and to pay attention to errors only when they cause a breakdown in communication.

Two Advanced prompts were used on Operational 1. The example papers cannot be reprinted here because of permission restrictions. In the first prompt, students were

asked to describe two photographs—in this case, a car and the Shinkansen or bullet train, and to tell how their trip would be similar or different depending on the mode of transportation they used.

For example, on one paper that received a score of 3, the highest score at the lower level, by applying the rubric, the raters would have considered whether there is enough development of ideas to put it into the 3 range. The sentences are in correct English word order, and there is some complexity in sentence structure as in the first sentence: "If we use the car to take a trip, a trip is good." The sentences and vocabulary are beginning to show some variety. We have the exclamation: Look at the picture 2! Mt. Fuji looks very beautiful. Vocabulary words are beginning to be varied as *in earth, gas, oil, beautiful.* There is even a growing sense of organization as we see the "first reason" that the train is good is that it is "better for the earth." The "second reason" is that "the view from the train is beautiful." There is even a concluding sentence, "So both the car and the train is good for taking a trip."

The second prompt for Operational 1, a persuasive or opinion prompt, asked students to agree or disagree on whether the school year in Japan should begin in September rather than in April.

In this second example, the paper received a score of 5, the next to the highest score on the rubric. The raters would have taken into account how clearly the thesis is expressed in the first line ("I think Japanese schools should begin in April, that is, I disagree with

this idea.") and how well the ideas are developed. Three reasons are given: in the first paragraph, that it is the traditional time to start school; in the second paragraph, that it's best to begin school when the cherry trees are blossoming; and, in the third paragraph, that it would take lots of hard work to change the starting time. Transitions are used to give an organizational pattern (First, Second, And Lastly). There is a concluding statement: "So, I disagree this idea." There is sentence variety and complexity. One interesting sentence is: "If Japanese school begin in September, we can't look that beautiful flower skin!!" There are two "if" clauses. Vocabulary that includes *rapidly, confused,* and *education* is rather advanced.

In addition to giving scores based on the rubric, scorers on this test also give feedback to the students in the form of coded comments. When Operational 1 was given, there were 56 comments divided into four areas: fluency and development, vocabulary, organization, and sentence elements. Each of the four areas included comments that were either Praise or Suggestions for Improvement. Essays were scored by two raters, with the first rater giving each essay four comments. The comments could come from any of the four areas that were most applicable to that paper. The one restriction was that each paper needed to receive at least one Praise comment and one Suggestion for Improvement. How the other two comments were divided depended on the paper itself. Giving individualized feedback on essays is rather unusual in the field of large-scale testing. Though the exact nature of the comments has changed somewhat, the Benesse English test continues to offer this personalized feedback to students.

III. The Data

The discussion of the data will focus on what relationships were found among scores on the different parts of the test, how these relationships compare with those found in previous studies involving other tests, and what the data may reveal about relationships among the skill areas of listening, reading, and writing.

The value of data on skills interrelationships

Data from the first operational administration of the Advanced version of the Proficiency Test of English Communication were analyzed to evaluate the performance of the instrument. In this report, the primary focus is on the observed relationships, or intercorrelations, among the skills of listening, reading and writing.

There are three main ways in which information gained from a proficiency test about skills interrelationships is important. First of all, what we can learn about the relationship among, or independence of, the language skills is of value in the development and verification of theories or models of language ability. Studies of intercorrelations among the skills have consistently shown that "there appears to be a significant amount of shared variance among the four skills" (Larson, 1983, p. 228).[1] The implications of this interrelationship among the macro-skills was a key issue in debates about the construct of language proficiency that came to a head with the provocative claims of Oller's (1979) Indivisibility Hypothesis (see, for example, Hosley & Meredith, 1979).

From these debates has emerged a consensus characterized as follows by Carroll (cited in Larson, 1983, p. 229): "There is a 'general language ability,' but at the same time ... language skills have some tendency to be developed and specialized to different degrees, or at different rates, so that different language skills can be separately recognized and measured."

The second way that evidence about skills interrelationships is important is in the practical task faced by schools and language institutes in deciding whether to use an additional test or test component—for example, a listening comprehension section or a writing task--for selection, placement, exiting or other purposes (see, for example, Hanania & Shikhani, 1986; Larson, 1983). Such decisions must weigh the value of unique information gained from a direct test of a particular skill against the additional burden—in terms of time, cost, and personnel—of a longer test or an additional component of a test battery.

Finally—and of greatest relevance to the test we are discussing in this paper--the measurement of skills intercorrelations is one of the most basic and widely used procedures for establishing the construct validity of a new test (see, for example, Educational Testing Service, 1997). Thus, for example, the Educational Testing Service (1989) asserted the construct validity of the Test of Written English (TWE) by arguing that "the degree of correlation between TWE scores and TOEFL scaled scores is low enough to allow for a conclusion that the TWE measures abilities distinct from, and in

addition to, those measured by TOEFL" (p. 12). Clearly, claims about the validity of the Benesse test will rest in part on data concerning skills interrelationships.

Previous research on second-language skills interrelationships

Figure 1 (see next page) lists the correlations between skills that have been found in selected studies. A few points are worthy of mention:

- There are relatively few comparisons of second-language reading and writing ability, and with one notable exception, these are all small-scale studies. This is not surprising in light of the fact that, until recently, the direct measurement of writing ability has rarely been a part of large-scale assessment of second-language proficiency.
- Findings do not always conform to intuitions and expectations about how the skills are related. For example, the correlation between second-language reading and writing, as measured by TWE and TOEFL Reading Comprehension scores, was in the overwhelming majority of comparisons (25 out of 30) lower than the correlation between TWE scores and TOEFL Listening Comprehension scores.

In general, research has consistently shown moderate to fairly strong interrelationships in performance in the different language skills. At the level of testing theory, these findings have led to the widespread view about the nature of second-language proficiency that although the different skills appear to tap a substantial core of knowledge and competences, each of the major skills is partially independent of the others, involving unique types of knowledge and distinctive processing or production operations.

FIGURE 1

Skills Intercorrelations in Previous Research

Listening-Reading

Educational Testing Service (1997)	.69 (TOEFL Listening/Reading)
Educational Testing Service (1991)	.84 (SLEP Listening/Reading)
Larson (1983)	.63 (2nd-year French examination) .86 (2nd-year German examination)

Listening-Writing

DeMauro (1992)	.57 (TOEFL Listening/TWE)
Educational Testing Service (1996)	.65 (5/95, Region 1, TOEFL Listening/TWE)
Educational Testing Service (1996)	.57 (5/95, Region 2, TOEFL Listening/TWE)
Educational Testing Service (1996)	.58 (5/95, Region 3, TOEFL Listening/TWE)

Reading-Writing

DeMauro (1992)	.54 (TOEFL Reading/TWE)
Educational Testing Service (1996)	.59 (5/95, Region 1, TOEFL Reading/TWE)
Educational Testing Service (1996)	.56 (5/95, Region 2, TOEFL Reading/TWE)
Educational Testing Service (1996)	.56 (5/95, Region 3, TOEFL Reading/TWE)
Carson et al. (1990)	.49 (for 48 Chinese students) .27 (for 57 Japanese students)
Hanania and Shikhani (1986)	.68 (cloze/written composition)

However, this uncontroversial position leaves much to be answered. Chief among these questions is how much independence of the skills one should expect to observe. For example, since the TOEIC and SLEP have produced strong correlations between listening and reading (.82 and .84, respectively), is the more moderate correlation between these skills on the TOEFL (.69) indicative of flaws in the way performance in these skills is being measured? Or is the TOEFL Listening/Reading correlation of .69 a more accurate index of the shared variance of these skills, and is it the much stronger correlations produced by the other two tests that should be questioned? All three tests make essentially the same argument: The observed correlations are all taken to indicate that the tests have construct validity. The weakness of such claims, however—and of the model of language ability on which they are based—is that it is difficult to refute them. Correlations as high as .90 (or higher) or as low as .50 (or lower) could presumably also serve as "evidence" of the partial independence of skills.

Statistical analysis of the Proficiency Test of English Communication

Descriptive statistics, calculated on the scaled scores, for the first operational administration of the test are reported in Table 1 (below)

TABLE 1

Descriptive Statistics for Scaled Scores on Benesse Test Sections

	Grade 9 (n=236)		Grade 10 (n = 10,142)		Grade 11 (n = 22,515)		Grade 12 (n = 2,071)		Total (N = 34,964)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Listening	16.57	3.37	16.83	3.64	16.40	3.57	18.05	4.26	16.63	3.66
Reading	15.28	4.28	15.39	4.43	15.82	4.54	18.65	5.56	15.86	4.63
Writing	7.60	2.67	7.92	2.53	7.48	2.72	8.13	3.21	7.65	2.71
Total	39.44	8.65	40.14	8.93	39.71	9.15	44.84	11.67	40.13	9.33

Skills intercorrelations

Skills intercorrelations are reported in Table 2 (below) and Figure 2 (next page).

TABLE 2

Intercorrelations of Scores on Benesse Subtests

	Grade 9	Grade 10	Grade 11	Grade 12	Total
Listening-Reading	.557	.620	.625	.724	.632
Listening-Writing	.483	.480	.479	.639	.495
Reading-Writing	.585	.520	.547	.702	.548

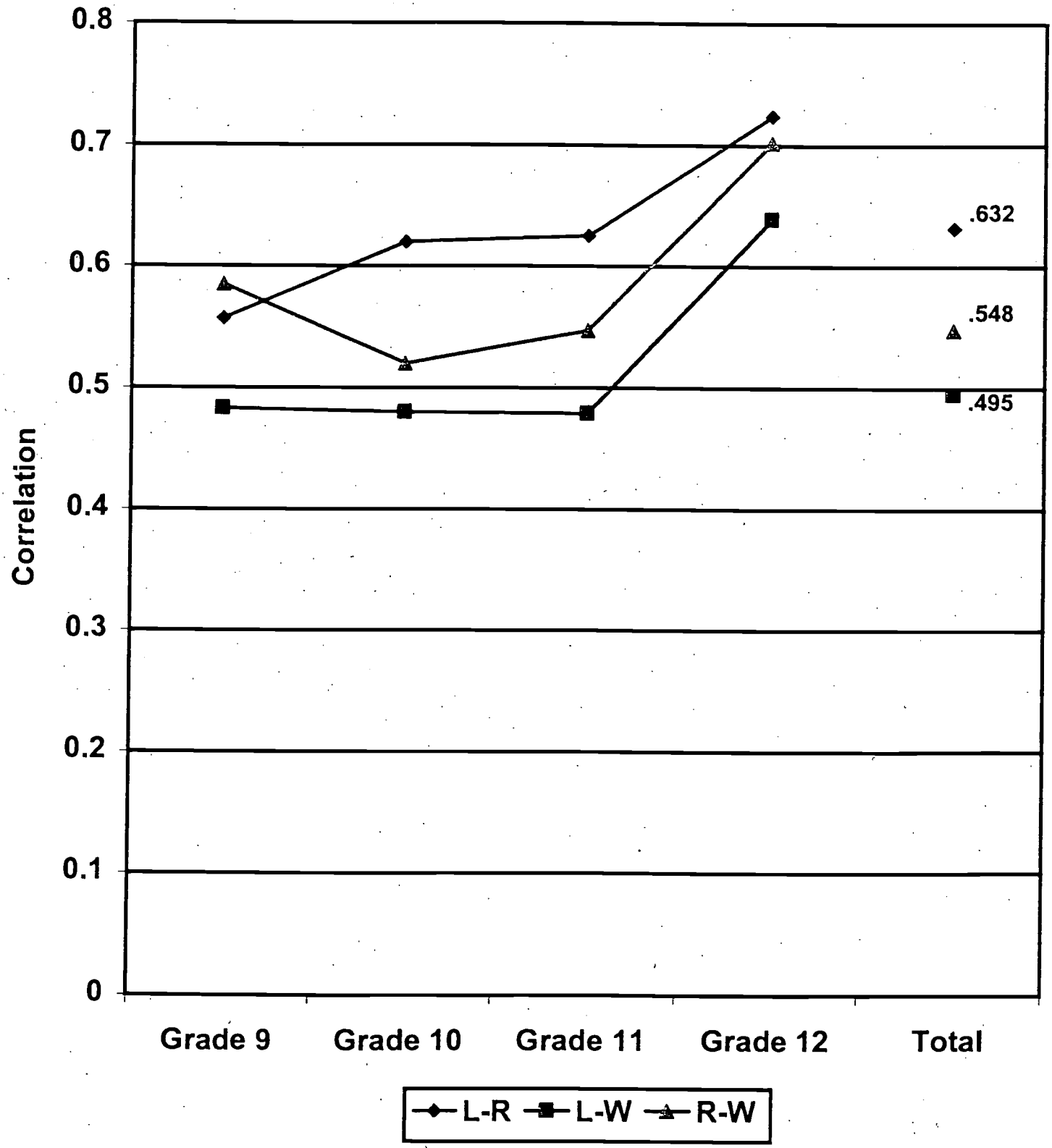
Notes: $N = 34{,}964$ (Grade 9 $n = 236$, Grade 10 $n = 10142$, Grade 11 $n = 22515$, Grade 12 $n = 2071$). All correlations are reported as Pearson product-moment correlation coefficients. All coefficients in Table 2 are significant at $p = .0001$.

The following general observations can be made:

- Generally, the highest correlations are between Listening and Reading; the next highest are the coefficients between reading and writing, and the lowest correlations are between listening and writing.
- With the exception of the correlations for Listening and Reading, which are similar to relationship of these skills on the TOEFL but considerably weaker than those produced by the TOEIC and SLEP, the coefficients are almost all very much in the range of those found in previous comparisons of performance in different skills (see Figure 1); the largest difference is between the correlations for Listening and Writing for Grades 9, 10, 11 and the Total sample and those reported by ETS (1996) in its comparison of TOEFL Listening and TWE scores—and even here, the coefficients would all be classified as "moderate."

FIGURE 2

Intercorrelations of Scaled Scores on Benesse Subtests



- The coefficients for listening-reading increase steadily Grade 9 through 12. However, for the other two skills intercorrelations, the coefficients remain fairly static from Grades 9 through 11 (with a slight decrease in listening-writing). Then for Grade 12 examinees the coefficients for all skill comparisons increase considerably—nearly .10 for listening- reading, nearly .16 for reading-writing, and a little over .16 for listening-writing. In addition to the increase in the magnitude of the coefficients for the Grade 12 students, the three coefficients for Grade 12 are more similar to one another than are the coefficients for examinees in Grades 9, 10 or 11.

At least two interpretations of the Grade 12 examinees' performance are possible: (a) There has been a significant change in their proficiency in one or more skill areas, or (b) they are different in some ways (type of school attended, academic aptitude, or any of a number of other potential variables) from the Grade 9, 10 and 11 examinees. Because information on background variables is not available, it is not possible to decide which interpretation is more plausible.

Regression analysis

A stepwise multiple regression analysis of the scaled scores indicates that an extremely large percentage of the variance in Total score is accounted for by the score on the Reading Section (see Table 3 below). Writing scores explain the smallest amount of

variance in Total scores. These findings are consistent with the skills intercorrelations observed.

TABLE 3

Summary of Stepwise Regression Procedure for Total (Scaled) Scores

	Partial r^2	Model r^2	$p =<$
Reading	.8153	.8153	.0001
Listening	.1290	.9442	.0001
Writing	.0558	1.0000	.1500

Note: All variables in the model are significant at the 0.15 level.

Reliability

Reliability[2] for the Listening subtest for this administration was .73; for the Reading subtest, .86; and for the Writing subtest, .78.

Conclusion

In this initial study of the first operational test results for the Benesse Proficiency Test of English Communication, it was found that the intercorrelations among the subtests are similar to those of other tests, which helps provide validation of this new test. The Benesse test was developed in response to the movement in Japan toward more communicative English language teaching and testing, and with this test, it is hoped that schools will be encouraged to change their English language curriculum and instruction to reflect a more practical and communicative approach.

[1] It must be kept in mind that strong intercorrelations between performances in different skill areas do not provide unambiguous evidence that they tap the same underlying trait; conversely, moderate and low correlations may result from factors other than the actual relationship between the skills.

[2] A number of British tests of proficiency in English as a second/foreign language have included separate measures of reading and writing. However, until recently, statistical analysis has not been a prominent part of test development by British examination boards (for a discussion of this issue, see, for example, Alderson, Slapha, & Wall, 1995; Alderson, Stansfield, & Krahnke, 1987).

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction. Cambridge: Cambridge University Press.

Alderson, J.C., Stansfield, C., & Krahnke, K. (1987). Reviews of English proficiency tests. Washington, DC: Teachers of English to Speakers of Other Languages.

Carson, J. E., Carrell, P. L., Silberstein, S., Kroll, B., & Kuehn, P. A. (1990). Reading-writing relationships in first and second language. *TESOL Quarterly, 24,* 245–266.

DeMauro, G. (1992). Examination of the relationships among TSE, TWE, and TOEFL scores. *Language Testing, 9,* 149–161.

Educational Testing Service. (1991). *SLEP test manual.* Princeton, NJ: Author.

Educational Testing Service. (1996). *Test of Written English guide.* Princeton, NJ: Author.

Educational Testing Service. (1997). *TOEFL test & score manual.* Princeton, NJ: Author.

Hanania, E, & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: Standardized ESL, cloze, and writing. *TESOL Quarterly, 20,* 97–109.

Hirai, A. (1999). The relationship between listening and reading rates of Japanese EFL learners. *The Modern Language Journal, 83,* 367–384.

Hosley, D., & Meredith, K. (1979). Inter- and intra-test correlates of the TOEFL. *TESOL Quarterly, 13,* 209–217.

Kitao, S. K., and Kitao, K. (1996). Testing communicative competence. *The Internet TESL Journal, Vol. II,* No. 5.

Larson, J. W. (1983). Skills correlations: A study of three final examinations. *Modern Language Journal, 67,* 228–234.

LoCastro, V. (1996). English language education in Japan. In H. Coleman, (Ed.), *Society and the Language Classroom* (pp. 40–58). Cambridge: Cambridge University Press.

Ministry of Education. (1989). *Course of study for junior high schools: Foreign languages—English.* Tokyo: Shoseki.

Ministry of Education. (1990). *Course of study for senior high schools: Foreign languages—English.* Tokyo: Shoseki.

Mulvey, B. (1999). A myth of influence: Japanese university entrance exams and their effect on junior and senior high school reading pedagogy. *JALT Journal, Vol. 21,* No. 1, 125–142.

National Institute on Student Achievement, Curriculum, and Assessment. (1998). *The educational system in Japan: Case study findings*. U.S. Department of Education: Office of Educational Research and Improvement.

Sato, K. and Kleinsasser, R. C. (1999). Communicative language teaching (CLT): Practical understandings. *The Modern Language Journal, 83*, 494–517.

FL 027201

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Interrelationships Among Skill Areas in Standardized Assessment	
Author(s): Stephen Gaies, Sandra Bolton, Diane Johnson, Caryl Lyons	
Corporate Source: TESOL 2001 Conference	Publication Date: Presented 3/1/01

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1 ↑ ☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑ ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑ ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Caryl Lyons	Printed Name/Position/Title: Caryl Lyons, Test Development Associate
Organization/Address: ACT, Inc, 2201 N. Dodge St., P.O. Box 168 Iowa City, IA 52243-1168	Telephone: (319) 337-1733 FAX:
	E-Mail Address: lyons@act.org Date: Feb. 19, 2002

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on
Languages & Linguistics
4646 40TH ST, NW
WASHINGTON, D.C. 20016-1859

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com