Analyzing an Achievement Test

到達度テストの検証

Hiroko Yoshida

本稿の目的は、ある大学の語学クラス（Academic Vocabulary Class）で使用された到達度テストを検証することである。到達度テストは目標基準試験（criterion- referenced test: CRT）であり、学習者の授業理解度を問うために、筆記試験として授業内で最もよく使用されている試験である。到達度テストは通常、成績に大きな割合を占めるが、試験実施後にそのテストを検証することはあまり行われていない。本研究では60項目からなる期末試験を項目分析（item analyses）を用いて検証した。まず、項目分析により、不良項目を削除した改訂版テスト（41項目版と27項目版）を作成した。さらに、オリジナルテストと２つの改訂版テストの項目基礎統計量（item statistics）、記述統計（descriptive statistics）、信頼性（reliability）、並びにファイ（ラムダ）指数（phi (lambda) dependability indexes）を分析した。その結果、14名の学生が受験したAcademic Vocabulary Classのオリジナルの期末試験は試験作成者の意図とは反して、約20％の項目が到達度テストとしての機能を十分に果たしていないことが示唆され、不良項目を削除した改訂版は到達度テストとして改良されたことが明らかになった。

An achievement test is the most relevant test for language teachers because it is probably the most frequently administered test in language programs. It occasionally plays an important part in evaluating student performance in the course or program and with the result that it would affect student motivation for subsequent learning. Furthermore, from the viewpoint of curriculum development, the results of the achievement test greatly affect curriculum evaluation if needs analysis is systematically administered (Brown, 1995). Therefore, the test should be fair whenever possible in every aspect: test questions, administration procedures, scoring methods and reporting policies (Brown, 1996). Nevertheless, evaluating achievement tests has been neglected in the language teaching context. The interests of most language teachers usually focus on making decisions of test content and methods in an achievement test. Once it is administered and scored, the test is rarely analyzed although it is an important part in meeting the teacher's demands for the development of sound classroom achievement tests. This study, then, aims to assess an achievement test actually conducted in the EFL classroom of Japan.

Literature Review

Language tests

In language courses or programs, different types of tests are used to make different types of decisions. Thus, selecting an appropriate type of test is imperative for the language teacher in making a given decision. The tests administered in language programs are basically categorized into four types: *proficiency tests, placement tests, diagnostic tests,* and *achievement tests.* They are used to make different types of decisions in language courses (Brown, 1995). A proficiency test is designed to assess how much of the language students know in order to make admission decisions. The focus of the proficiency test is to evaluate general, overall language ability without reference to any particular program. A placement test examines general knowledge of language as the proficiency test does; however, it differs from the proficiency test in that the placement test assesses the relatively narrow range of abilities for a given program and it aims to stream students into different levels within the program. Proficiency tests and placement tests are both norm-referenced tests (NRTs) which are designed to measure comprehensive language abilities. Each student's score on NRTs is interpreted with reference to the scores of all other students who participated in the test. The dispersion of scores in NRTs usually depicts normal distribution. Students generally have little knowledge of questions in NRTs, although they may be familiar with question formats (Brown, 1989, 1995, 1996). On the other hand, a diagnostic test assesses the degree to which the specific instructional goals of the course or program have been accomplished in a given class. It is commonly administered at the beginning or in the middle of the language course. An achievement test is also designed to assess the extent to which students have mastered course objectives, but it is commonly conducted at the end of a course or program. A diagnostic test and an achievement test are called criterion-referenced tests (CRTs). Language teachers should bear in mind that the features of CRTs totally differ from those of NRTs; CRTs aim to examine the extent to which specific instructional objectives have been achieved by each student. They are designed to compare a student's performance with, not the other students' scores, but only particular learning objectives of the course or program (Brown, 1996). On CRTs, the students normally know not only what item types to be expected in the test, but also what language points to be tested before they actually take the test if objectives are clearly stated and they are well instructed (Brown, 1996). In an achievement test, it is not rare that study questions are given to the students before the implementation of the test to help students review and prepare for the test.

Brown (1996) claims that understanding the differences between CRTs and NRTs leads to

making better decisions about students and developing and analyzing the tests. However, the distinction has not been sufficiently recognized by many language teachers although it has been discussed in the language testing literature (Bachman, 1989; Brown, 1989, 1990, 1993). The different test qualities that these four tests have, i.e., detail of information, purpose of decision, relationship to program, administration timing, and interpretation of scores are shown in Table 1.

Table 1
Tests Qualities of Four Tests

	Types of Decision			
	Norm-Referenced		Criterion-Referenced	
Test qualities	**Proficiency**	**Placement**	**Diagnostic**	**Achievement**
Detail of Information	Very General	General	Very Specific	Specific
Focus	Usually, general skills prerequisite to entry	Learning points all levels and skills of program	Terminal and enabling objectives of courses	Terminal objectives of course or program
Purpose of Decision	To compare individual overall with other groups/individuals	To find each student's appropriate level	To inform students and teachers of objectives needing more work	To determine the degree of learning for advancement of graduation
Relationship to Program	Comparisons with other institutions	Comparisons within programs	Directly related to objectives still needing work	Directly related to objectives of program
Administration Timing	Before entry and sometimes at exit	Beginning of programs	Beginning and /or middle of courses	End of courses
Interpretation of Scores	Spread of scores	Spread of scores	Number and amount of objectives learned	Number and amount of objectives learned

Note: From *Testing in Language Program.* (p.9) by J. D. Brown, 1996, NJ: Prentice Hall Regents. Copyright 1996 by Prentice Hall Regents. Adapted with permission of the author.

Assessments

Assessing language knowledge consistently is not simple; any test cannot be immune from a certain amount of errors (Brown, 1996). Nevertheless, Brown (1996) insisted that language testers should be concerned with its consistency whenever possible. To this end, he used statistical analyses and examined test consistency (Brown, 1989, 1990, 1993). Item analysis is designed to examine the degree to which the individual items on a test are effective. Three statistical analyses are used to analyze items of a test: item facility analysis, B-index analysis, and item discrimination analysis. Analyzing the items on CRTs enables teachers to make decisions about which items are to be kept and which items are to be deleted (Brown, 1996).

Item facility (IF) shows the proportion of students who answered a given item correctly (Brown, 1995). This index is calculated by adding the number of students who correctly answered an item and dividing that sum by the total number of students who took the test. The

yielded result is an index ranging from 0.00 to 1.00. The index would be the percentage of correct answers for a particular item. For example, an IF index of .70 can be interpreted as 70% of the students correctly answering the item. This item is regarded as a relatively easy question. On the other hand, an item with an IF of .15 would be a difficult question because 85% of the students incorrectly answered the item (Brown, 1996).

The B-index is the difference between proportions of correct answers on each item and the proportions of students passing and falling (Brown, 1993, 1996). It shows the degree to which the students who passed the test outperformed the students who failed the test on each item. The B-index firstly determines the cut-point for passing the test and then compares the IFs of those students who passed a test with the IFs of those who failed it. For example, if the cut-point of 70% is determined, "students who passed the test" means students who answered correctly 70% or more of the items, while "students who failed the test" means students who answered correctly below 70% of the items. The IF indexes are next to be calculated for two groups: item facility for students who passed the test and item facility for students who failed the test. The B-index is represented as the difference between two item facility indexes. For example, when IF in the pass group is 1.00 in a particular item (i.e., 100% correctly answered the item in the pass group) and IF in the fail group is 0.00 (0% correctly answered the item in the fail group), the B-index is 1.00 (1.00 - 0.00 = 1.00). This shows that the given item sufficiently distinguishes between students who passed the test and students who failed the test. The resulting B-index values can range from -1.00 to +1.00.

Item discrimination (ID) is an index of the degree to which a given item separates the upper third of the students from the lower third of the students (Brown, 1996). It is designed to compare the performance of the high-scored students on the test with that of the low-scored students. To calculate ID, the IFs for the upper and lower groups for each item are respectively determined, and the IF for the lower group is subtracted from the IF for the upper group. The resulting ID value can range from -1.00 to +1.00. When all of the students in the lower group correctly answer and those who in the upper group incorrectly answer, the ID would be -1.00, whereas when all students in the high-scored students correctly answer and those who in the lower incorrectly answer, it would be +1.00. Brown (1989, 1996) introduced guidelines to judge items based on ID as follows, by citing Ebel (1979).

.40 and up	Very good items
.30 to .39	Reasonably good but possibly subject to improvement
.20 to .29	Marginal items that are usually subject to improvement

Below .19　　Poor items that are to be deleted or improved by revision

Another important element of the test is reliability, which means that a test yields the identical or very similar results whenever it is conducted under the same conditions. Producing consistent results in a test if the students were to take it repeatedly is desirable in any measurement regardless of whether it is norm-referenced or criterion-referenced. Internal-consistency reliability is used to estimate reliability when a single NRT is administered only once. Examples are alpha coefficient, the Kuder-Richardson formula 21 (K-R21) and the Kuder-Richardson formula 20 (K-R20), which are known as relatively easy procedures to calculate internal consistency (Brown, 1996). On CRTs, threshold loss agreement, squared-error loss agreement, and domain score dependability are employed to measure reliability[1)] (Brown, 1996). Brown (1990) examined criterion-referenced test reliability by using these three approaches. Since explaining the details of all measurements of reliability is beyond the scope of this paper, only the phi (lambda) dependability index, which is one of squared-error loss agreement approaches, is presented here. It can estimate reliability in a CRT which is administrated once and attempts to account for the distances that students are from the cut-point for the master/non-master classification. The yielded index ranges from 0.00 to 1.00. For example, a phi (lambda) dependability index of .90 suggests that the test is highly reliable.

Although many language testers acknowledge the importance of examining language tests, few attempts have been made to investigate an achievement test used in the classroom in Japan. The purpose of this study, then, is to examine an achievement test actually conducted in the EFL classroom. To this end, the following questions were posed:

1. What are the item statistics for the original and revised versions of a criterion-referenced vocabulary test?
2. What are the descriptive statistics for the original and revised versions of the program-related vocabulary test?
3. To what degree are the original and revised versions of the test reliable?
4. To what degree are the phi (lambda) dependability indexes consistent with different cut-points?

Methods

Participants

Participants for this study consisted of 14 college students whose first language (L1) was Japanese. They enrolled in an academic vocabulary class, which was an elective course taught by

the author at a college in a western part of Japan. All participants were female and their age ranged from 18 to 21. According to an in-house placement test, their language proficiency was at the intermediate level.[2)]

Material

The material used in this study was a vocabulary test that was conducted at the end of the course as a final examination. It consisted of three sections: a fill-in test (Section I), a translation test (Section II), and a test based on a worksheet (Section III). The fill-in test was given together with a list of options. The students were familiar with this part because they had had quizzes twice using the same procedure during the course before the achievement test. In the translation test, the students were required to translate given Japanese words into English. The students had been given study questions beforehand and all items in this section came from the study questions. The third section employed the same questions as they were introduced in a worksheet actually used in the class. The original version of the vocabulary test consisted of 60 items: 30 items for the fill-in test, 25 items for the translation test, and 5 items for the worksheet test (Appendix).

Procedures

The test was administered to 14 students at the end of the course in the classroom. The students were given 50 minutes to finish the 60 items. All the items were scored in the same procedure; right answers were counted as one point each, while wrong answers received no points. Thus, the perfect score for the test was 60 points.

Analysis

The data obtained in the vocabulary test was examined in terms of descriptive statistics, which include the number of items (k), number of participants (N), mean (M), standard deviation (SD), and range. Two reliability estimates were also calculated: the Kuder-Richardson formula 21 (K-R 21) and the phi (lambda) dependability index (Φ). Although the phi (lambda) dependability was used to examine reliability, these agreement coefficients are dependent on the cut-point, which has been occasionally criticized. To deal with this problem, this study set three cut-points (90%, 80%, and 70%) and compared the results. The 60 items were then analyzed individually based on item facility, B-index, and item discrimination to choose the items for the revised versions. In selecting items from the original test, two criteria were employed. In the first revised test, items that fell approximately within a range of .25 to 1.00 in B-index and had an item

discrimination near or in excess of .20 were kept. As a result, the number of items kept was 41 (Revised 41). The second revised test, only those that fell approximately within a range of .30 to .80 in B-index and had an item discrimination near or in excess of .30 were kept, and consequently only 27 items were selected (Revised 27). Furthermore, these revised versions were then analyzed for descriptive statistics and item analysis to examine the degree to which the revisions succeeded.

Results

The decisions about which items to keep in the revised versions and which items to discard were based on the results of item facility, B-index, and item discrimination shown in Table 2.

Table 2
Item Analyses of the Original Test

Item Number	Item Facility	B-Index	ID	Item Number	Item Facility	B-Index	ID
*1	1.00	0.00	0.00	+31	0.93	0.25	0.20
*2	1.00	0.00	0.00	+32	0.93	0.25	0.20
*3	1.00	0.00	0.00	*33	0.79	0.05	0.00
+4	0.93	0.25	0.20	34	0.86	0.50	0.40
5	0.79	0.40	0.20	+35	0.93	0.25	0.20
6	0.79	0.75	0.60	+36	0.93	0.25	0.20
7	0.50	0.35	0.20	*37	1.00	0.00	0.00
*8	1.00	0.00	0.00	38	0.79	0.75	0.60
9	0.79	0.75	0.60	+39	0.93	0.25	0.20
10	0.79	0.40	0.40	40	0.64	0.90	0.80
*11	1.00	0.00	0.00	41	0.79	0.75	0.60
*12	1.00	0.00	0.00	*42	1.00	0.00	0.00
+13	0.93	0.25	0.20	+43	0.93	0.25	0.20
+14	0.93	0.25	0.20	44	0.79	0.75	0.60
+15	0.93	0.25	0.20	*45	0.86	0.15	0.20
*16	1.00	0.00	0.00	46	0.64	0.55	0.80
*17	1.00	0.00	0.00	47	0.71	1.00	0.80
18	0.86	0.50	0.40	+48	0.93	0.25	0.20
+19	0.93	0.25	0.20	49	0.86	0.50	0.40
20	0.86	0.50	0.40	+50	0.93	0.25	0.20
21	0.71	1.00	0.80	51	0.50	0.70	0.80
22	0.79	0.75	0.60	52	0.71	1.00	0.80
23	0.71	0.30	0.60	+53	0.93	0.25	0.20
*24	1.00	0.00	0.00	54	0.71	1.00	0.80
25	0.64	0.90	1.00	55	0.43	0.60	1.00
26	0.64	0.55	0.80	*56	0.93	−0.10	0.20
*27	1.00	0.00	0.00	*57	0.64	−0.15	0.20
28	0.71	1.00	0.80	*58	0.64	−0.15	0.40
29	0.79	0.75	0.60	*59	0.93	−0.10	0.20
30	0.79	0.75	0.60	*60	0.93	−0.10	0.00

Note: Items with an asterisk (*) were not included in the Revised 41 version and items with a plus (+) and an asterisk (*) were not included in the Revised 27 version.

Items with an asterisk (*) were not included in the Revised 41 version and items with a plus (+) and an asterisk (*) were not included in the Revised 27 version. In Revised 41, most of the selected items had B-indexes between 0.25 and 1.00 and most of the selected items had IDs near or in excess of .20. In Revised 27, most of the selected items had a B-index between .30 and .80 and most of the selected items had an ID near or in excess of .30. After the items were deleted, the results of the achievement test were reanalyzed as if the 41 and 27 selecting items

Table 3
Item Analyses of the Revised 41 Test

Item Number	Item Facility	B-Index	ID	Item Number	Item Facility	B-Index	ID
4	0.93	0.25	0.20	32	0.93	0.25	0.20
5	0.79	0.40	0.40	34	0.86	0.50	0.40
6	0.79	0.75	0.60	35	0.93	0.25	0.20
7	0.50	0.35	0.20	36	0.93	0.25	0.20
9	0.79	0.75	0.60	38	0.79	0.75	0.60
10	0.79	0.40	0.40	39	0.93	0.25	0.20
13	0.93	0.25	0.20	40	0.64	0.90	0.80
14	0.93	0.25	0.20	41	0.79	0.75	0.60
15	0.93	0.25	0.20	43	0.93	0.25	0.20
18	0.86	0.50	0.40	44	0.79	0.75	0.60
19	0.93	0.25	0.20	46	0.64	0.55	0.80
20	0.86	0.50	0.40	47	0.71	1.00	0.80
21	0.71	1.00	0.80	48	0.93	0.25	0.20
22	0.79	0.75	0.60	49	0.86	0.50	0.40
23	0.71	0.30	0.60	50	0.93	0.25	0.20
25	0.64	0.90	1.00	51	0.50	0.70	1.00
26	0.64	0.55	0.80	52	0.71	1.00	0.80
28	0.71	1.00	0.80	53	0.93	0.25	0.20
29	0.79	0.75	0.60	54	0.71	1.00	0.80
30	0.79	0.75	0.60	55	0.43	0.60	1.00
31	0.93	0.25	0.20				

Table 4
Item Analyses of the Revised 27 Test

Item Number	Item Facility	B-Index	ID	Item Number	Item Facility	B-Index	ID
5	0.79	0.21	0.40	30	0.79	0.50	0.60
6	0.79	0.50	0.60	34	0.86	0.33	0.40
7	0.50	0.29	0.20	38	0.79	0.50	0.60
9	0.79	0.50	0.60	40	0.64	0.54	0.80
10	0.79	0.50	0.40	41	0.79	0.50	0.60
18	0.86	0.33	0.40	44	0.79	0.50	0.60
20	0.86	0.33	0.40	46	0.64	0.54	0.80
21	0.71	0.67	0.80	47	0.71	0.67	0.80
22	0.79	0.50	0.60	49	0.86	0.33	0.40
23	0.71	0.67	0.60	51	0.50	0.88	1.00
25	0.64	0.83	1.00	52	0.71	0.67	0.80
26	0.64	0.83	0.80	54	0.71	0.67	0.80
28	0.71	0.67	0.80	55	0.43	0.75	1.00
29	0.79	0.50	0.60				

had been administered. The new item statistics were reported in Table 3 and 4 respectively. This analysis roughly estimated what would happen if we used these two revised versions.

The descriptive statistics for the original test, the Revised 41, and the Revised 27 are reported in Table 5. Phi (lambda) dependability indexes were analyzed according to three different cut-points (90%, 80%, and 70%) of the original test, the Revised 41, and the Revised 27 (Table 6).

Table 5
Descriptive Statistics

Statistics	Original test	Revised 41	Revised 27
k	60.00	41.00	27.00
M	50.29	32.57	19.57
SD	10.76	10.74	8.99
Range	30.00	29.00	22.00
K-R21	.25	.39	.42

N = 14

Table 6
Phi (lambda) Dependability Index

Cut-point	Φ (.90)	Φ (.80)	Φ (.70)
Original test	.95	.95	.97
Revised 41	.97	.97	.97
Revised 27	.98	.97	.97

Discussion

The item statistics for the original vocabulary achievement test clearly indicated that almost 20% of the items in the original version were not appropriate for the test. Especially, the items in Section III, which were based on a worksheet actually used in the classroom showed that it did not function at all. This was a totally unexpected result, as the students were expected to be familiar with these questions. Although the questions in Section III were designed to make the students review worksheets used in the classroom, the results suggested that the teacher's intention did not work efficiently as initially expected.

As for the second research question, the descriptive statistics indicated that the Revised 27 would function most effectively as an achievement test as it produced the lowest standard deviation (*SD*). As CRTs are not designed to produce variance in scores, producing little variance in a CRT indicated that the test appropriately functioned as a sound CRT (Brown, 1990, 1996). The KR-21 indicated that the two revised tests were slightly more reliable than the original test, but none of the estimates were considerably high. On the other hand, despite the different cut-

points, the differences in the phi (lambda) dependability indexes of the original test, the Revised 41 and the Revised 27 were not striking, and all of them were high. It indicated that the scores of all three tests were considered reliable.

The differences between the K-R 21 and the phi (lambda) dependability indexes may result from the score distribution of the vocabulary tests; they were negatively skewed and did not show normal distribution. When the standard deviation goes down relative to other factors, such as the number of items, and the mean of the test scores, the internal-consistency will decrease as the Kuder-Richardson formula 21 is sensitive to the degree of the standard deviation. Thus, in CRTs, which are not designed to produce variance in scores, phi (lambda) dependability indexes are more reliable than K-R 21 (Brown, 1996).

In short, the original test as well as the Revised 41 and the Revised 27 is highly consistent and reliable. Furthermore, the difference of the cut-point did not affect the degree of consistency. Phi (lambda) dependability indexes of three levels of the cut-point of the original test, the Revised 41, and the Revised 27 were all consistently high.

Conclusion

This study has evaluated a program-related vocabulary test. Despite its high reliability, analyzing test items of the original test revealed that some 20 % of the questions were not appropriate to evaluate students' learning as an achievement test. Based on these outcomes of item analysis, two revised tests were formed and reanalyzed. The results indicated that the revised versions are more preferable than the original test and have slightly higher reliability. Item analysis successfully improved the program-related achievement test in which the test maker's intention did not function in some items, despite high reliability of the original test. The results suggested that the original test needed to be revised.

Language tests play multiple important roles in language curriculum. For students who invest a great amount of time and energy in learning the language, the test is expected to meet their demands. Students who have made efforts to prepare for it should obtain higher scores than others who have not in an achievement test. The test items should be fair enough to reflect objectives and goals of the course. For language teachers, developing sound CRTs affects the cyclical process of the curriculum. Examining the test that was actually used in the classroom leads to an effective revision of materials and teaching (Brown 1993). Given the significant effects that the test poses, it is highly desirable that achievement tests are examined after they are scored and reported based, not only on test makers' intuitions, but also on objective analyses.

Lastly, the paper did not refer to validity and usability; however they are also important components to be considered in testing (Brown 1996). Validity means the extent to which a test measures what it is supposed to measure, whereas usability concerns the extent to which a test is practical to actually implement. They are quite different test characteristics; however, they are all necessary in sound CRTs. Language testers should also keep in mind validity and usability in assessing an achievement test.

Notes

1）Brown (1996) differentiated terms to express consistency of the different types of tests; reliability is used for NRTs, while dependability is used for estimates of the consistency of CRTs so as to understand the differences between the notions of NRTs and CRTs. However, in this paper, the term, reliability, is used for expressing consistency of both NRTs and CRTs.

2）Consent to release the details of the students' English proficiency was not obtained.

References

Bachman, L. F. (1989). The development and use of criterion-referenced tests of language proficiency in language program evaluation. In K. Johnson, (Ed.), *Program design and evaluation in language teaching*. Cambridge: Cambridge University Press.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23* (1), 65-83.

Brown, J. D. (1990). Short-cut estimators of criterion-referenced test consistency. *Language Testing, 7* (1), 77-97.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In C. C. D. Douglas (Ed.), *A new decade of language testing research* (pp. 163-184). Washington, D.C.: TESOL.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. New York: Heinle & Heinle Publishers.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice-Hall.

Ebel, R. L. (1979). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Ito, S. (1995). *Kokusai nyusu o kaku eigo (English for international news)*. Tokyo: Chigasaki Syuppan.

Appendix

Original Test (Answer sheet omitted)

Ⅰ．次の空所に当てはまるものをあとの語群から選んで答えなさい。ただし、文法的に正しくなるように適当な形に変化させること。（２回使用する語もあり）

1．税務署は適当な領収書がなければ払い戻し申請を**受理**しない。

The taxation office will not（　1　）your request for a tax rebate if you don't have the proper receipts.

2．西側諸国の中には統一ドイツの影響力を**恐れる**国もあった。

Some Western countries（　2　）the influence of a unified Germany.

3．日本はポーランドなど東欧諸国との友好を促進する**用意がある**。

Japan（　3　）promote friendship with Poland and other East European countries.

4．（国民）大衆のその政党に対する不信が選挙での当の敗北を**もたらし**たに違いない。

The public distrust of the party must have（　4　）its defeat in the election.

5．日本は世界の全ての地域で自由貿易を促進**する立場にいる**。

Japan（　5　）promoting Free Trade in all parts of the world.

6．新大統領は経済政策に対して高まる批判に**対応し**なければならない。

The new President will have to（　6　）mounting criticism of his economic policy.

7．最近の決定によってその国の経済刺激策が**用意され**た。

The recent decision（　7　）measures to stimulate the nation's economy.

8．その元大統領は賄賂を受け取った**というかどで**起訴された。

The former President was indicted（　8　）receiving bribes.

9．その国の労働者は賃金引上げを要求**せずに**働くことを要請された。

Workers of the country were requested to work（　9　）demanding higher wages.

10．彼はその国の指導者に対し民主化へ更に努力するよう**説得する**ことを強く要請された。

He was urged to（　10　）the leaders of the country to do more for democratization.

11．核兵器の**削減**は世界平和への道である。

（　11　）nuclear armaments would lead to world peace.

12．彼の抜本的な改革政策は**結果として**ソ連共産党の解体となった。

His drastic reform policy（　12　）the disbandment of the Soviet Communist Party.

13．教科書問題で中国人は彼らの国で日本の過去の行為を**思い出した**。

The textbook issue（　13　）Chinese people（　13　）Japan's past conduct in their country.

14．日本は非核三原則を**堅持する**と約束している。

Japan has promised that it will（ 14 ）its three non-nuclear principles.

15. 新指導部は国の経済を安定させることを**強く求められて**いる。

The new leadership（ 15 ）stabilize the nation's economy.

16. 民間の調査によればアメリカの若者は日本の若者より政治に対する**満足度**が強い。

A private survey shows that American youngsters（ 16 ）more（ 16 ）politics than their Japanese counterparts.

17. 日本は選挙後、政治危機に**直面**した。

Japan（ 17 ）a political crisis after the election.

18. 日本は議会制民主主義を確実にするため金権政治を**脱し**なければならない。

Japan should（ 18 ）money-powered politics to ensure parliamentary democracy.

19. 同党は都議会議員選挙で勝利に**向かって**いる。

The party is（ 19 ）a victory in the metropolitan assembly election.

20. 政治改革についての勧告は年末までに**提出**される。

A recommendation on political reform will（ 20 ）by the end of the year.

21. 政府はその財界人の国家に対する貢献を高く**評価した**。

The government highly（ 21 ）the businessman's contribution to the nation.

22. 発展途上国は、生活水準で先進工業国にはるかに**遅れている**。

Developing countries are（ 22 ）far（ 22 ）industrial countries in their standard of living.

23. たいていの経済学者たちは、ここ数年間はなだらかな経済成長が続くと**予測している**。

Most economists（ 23 ）moderate economic growth in the years to come.

24. 内外の需要に応えるため、工場の操業は**全開状態**である。

Operations at the factory（ 24 ）to meet the growing demand at home and abroad.

25. 父はもうこれ以上私たちにお金を**くれない**と言っている。

Father tells us he will not（ 25 ）us with money any more.

26. 日本の対米貿易黒字は400億ドル以上と**推定されている**。

Japan's trade surplus with the United States（ 26 ）more than 40-billion dollars.

27. 東京の北方の山中に飛行機が墜落し、**少なくとも**500人が死亡した。

（ 27 ）500 people were killed when a plane crashed into a mountain north of Tokyo.

28. 現在のところ他の詳細は**わからない**。

No other details（ 28 ）at present.

29. 日本人の平均寿命はまだ伸び続けるものと**みられる**。

In, Japan, the average life expectancy（　29　）be further extended.

30. 新会社の発足の祝賀行事が**行われて**いる。

Festivities（　30　）, celebrating the inauguration of the new company.

avoid, at least, agree, admit, accept, adopt, appoint, appreciate, at least, be banned, be likely to, be anxious about, be available, bring about, be concerned over, be satisfied with, be under way, be ready to, be in full swing, be committed to, be afraid of, be at a loss, call on ~to, conclude, cope with, confer, do without, estimate at, face with, facilitate, get rid of, give rise to, head for, in line with, in accordance with, instead of, lead to, lag behind, mainly, manage to, on charges of, play an important role, persuade, pay, predict, provide, result in, remind of, reduce, stress, stem from, stick to, submit, share with, urge to, warn,

Ⅱ．次の言葉を英語に直しなさい。

1．経済改革
2．週休二日制
3．首相
4．駐日大使
5．アメリカ政府
6．２カ国の(両国間の)関係
7．発展途上国
8．過去の侵略行為
9．近隣諸国
10. 国際社会
11. 議会制民主主義
12. 世界平和
13. 反日感情
14. 内閣
15. 熱帯雨林の破壊
16. 代表団
17. 総選挙
18. 地滑り的勝利
19. （日本の）国会
20. 参議院
21. 日本国憲法
22. 個人消費の停滞
23. 生活水準
24. 貿易不均衡
25. 円

Ⅲ．下線部の単語の中に含まれている単語を見つけたうえで、次の文を日本語にしなさい.

(ex.)　Mr. X is an <u>indecisive</u> leader.

単語　decide

意味　Xさんは優柔不断なリーダーだ。

1．I am very happy to accept your *<u>invitation.</u>*

2 . It is an *exclusive* interview.

3 . Teenagers are highly *suggestible*.

4 . The pandas are *endangered* species.

5 . You should have an *animated* discussion.

Note. The questions of Sections I and II are based on the course textbook, *Kokusai nyusu o kaku eigo* (Ito, 1995).