Item Analyses of a Multiple-choice Achievement Test: Increasing its Usefulness

Kaori Nitta

1. Introduction

Like many other universities, Kinki University has tried to have our students get higher scores on the TOEIC Test. We changed our syllabus to realize a communicative-oriented course, wrote a textbook focusing on preparing for the TOEIC Test, and began to implement an achievement test that would be also used as a screening test for the real TOEIC Test, which is a proficiency test.

In this paper, I will state that it is possible to use a test for multiple purposes, such as an achievement test as a proficiency test. Next, I will suggest that we use two kinds of item analyses to make our test more useful. Finally, I will suggest that for a multiple choice test, we use two distractors instead of three.

2. An achievement test and a proficiency test

Is it appropriate to use one test for multiple purposes? First of all, what is an 'achievement test' ? What is the difference between an 'achievement test' and a 'proficiency test' like TOEIC or TOEFL? Alderson, Clapham and Wall (1995: 286) say that achievement tests are "similar to progress tests, but they are given at the end of the course. The content ... [are] generally based on the course syllabus or the course textbook." On the other hand, they claim that proficiency tests are "not based on a particular language programme ..." and "show whether students have sufficient ability to be able to use a language in some specific area ..." (293) In other words, achievement tests are used to judge the past, namely, whether students

have achieved a certain level at the end of the course. On the other hand, proficiency tests are the ones to predict the future, that is, whether they can survive successfully in their future discourse community, such as graduate schools and companies they will belong to. Thus, two types of tests apparently have different purposes.

However, according to Brindley (1986), both types of tests are blurred, and hard to distinguish in a communicatively-orientated language program. Hughes (2003) also states that both kinds of tests are essentially the same. Moreover, McNamara (1996) claims that if the course syllabus reflects the real world, one test can be used for different purposes. So a test can have multiple purposes under certain circumstances such as in a communicatively-oriented foreign language course.

In Kinki University, our English course for first-year students is intended to help them prepare for the TOEIC Test. This test is based on business communication in the real world, that is, communicative English skills and knowledge companies expect their employees to have in order to make their business work well in the global society. So our English course is also connected to the real world outside the campus.

Our goal is to help our students improve their communicative skills, which, eventually, enables them to get higher scores on the TOEIC Test. However, questions of real TOEIC Tests are too difficult for most of our students, who need step-by-step exercises before challenging the real TOEIC questions. Since we could not find any suitable textbook including such exercises, we wrote a new textbook for the course, focusing on the basic communicative skills.

By analyzing the questions of the TOEIC Test, we can decide which functions, which grammatical points, and what kinds of genres for reading our students should acquire. We included what we thought was the most important on the basis of the priority in the textbook, and so the textbook is itself concerned with the TOEIC Test, that is, the real business-oriented domain of the outside world.

Our achievement test for this course is used to check how much our students have acquired through the textbook, as Nunan (1999: 301) claims that an achievement test "attempts to measure what students have learned from a particular course or set of materials." At the same time, it is also possible to predict students' performances on the real TOEIC Test, because our course and the TOEIC Test include the same functions, the same grammatical points and the same genres for reading. Thus, since the course syllabus reflects the real world, our test can be used for different purposes: an achievement test and a screening test for the real TOEIC Test.

3. Item analyses

3.1 Why is statistical analysis needed

It is, of course, impossible to develop a perfect test, in particular, which has multiple purposes. However, we can make efforts to make our tests more useful, or as useful as possible. In order to justify the test, we need to have some evidence. Bachman and Palmer (1996: 17) states that "the most important quality of a test is its usefulness. --- We thus regard a model of test usefulness as the essential basis for quality control throughout the entire test development process. --- we propose a model of test usefulness that includes six test qualities-reliability, construct validity, authenticity, interactiveness, impact, and practicality."

We use a multiple-choice test as well as the TOEIC Test. Since we want to help students get higher scores on the TOEIC Test, to give a multiple-choice questions is authentic in a sense. Moreover, we use as many authentic materials as possible.

It goes without saying that multiple-choice tests are very practical, that is, easy to implement and rate. There is no inconsistency among raters. They, however, tend to lack interactiveness because those tests only give students limited tasks.

We may give a positive washback (impact) to all the classes by giving the same test to the

students of these classes in order to evaluate their class performances. Of course, we need to be very careful not to make our students too test-wise.

As I said above, our achievement test is supposed to measure communicative English skills and knowledge companies expect their employees to have. We interpret test takers' performances on the test, that is, to what extent test takers have obtained. To investigate the scores test takers obtain definitely includes quantitative statistical procedures which are related to two of the qualities of measurement, reliability and construct validity.

According to Bachman and Palmer (1996: 19), reliability is "defined as consistency of measurement." And construct validity is "the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores." (21)

In order to justify the interpretations, Bachman and Palmer (1996: 21) claim that "we need to provide evidence that the test score reflects the area(s) of language ability we want to measure." Bachman (2004: 3) points out that "we need to be able to demonstrate that scores we obtain from language tests are reliable, and that the ways in which we interpret and use language test scores are valid. --- An important kind of evidence --- is that which we derive from quantitative data – scores from test tasks and tests as a whole – and the appropriate statistical analyses of these data."

In the next section, I will focus on how to increase reliability of our test by using two kinds of item analyses: classic item analyses and IRT (item response theory). IRT has been developed to solve the problems of classic item analyses, but of course, it has its problems or limitations, so we should know what can be more suitable to analyze test items. To increase reliability of our test will lead us to better validity, and reliability is one of the tools to make a test more useful.

3.2 Analyses of our achievement test

3.2.1 What is the achievement test like

The test which is analyzed here is the achievement test administered in November, 2004. The purposes of the test are:

1) to check how much the students have mastered what they have been exposed to
2) to select the upper 40 % of the students who can take TOEIC for free

The details are as follows:

◆ The participants: 236 first-year students who are non-English majors
◆ The number of the items: 75 items (Listening section; 30, Reading section: 45)
◆ The time duration: 60 minutes
◆ The content: 1) Listening section: 30 items
(1) photos: 8 items (4 options)
(2) quick responses: 10 items (3 options)
(3) conversations: 7 items (4 options)
(4) announcements: 5 items (4 options)
2) Reading section: 45 items
(5) incomplete sentences: 20 items (4 options)
(6) error recognition: 10 items (4 options)
(7) reading comprehension: 15 items (4 options)

Like the case of the TOEIC Test, our achievement test has seven sections as above, but the numbers of each section are different from the TOEIC Test because we chose more suitable questions for 1-year students depending on the difficulty of each section.

3.2.2 Procedures of analyzing the achievement test

Before we implemented the achievement test, we had given a pilot test to about 100 second-

year students, and modified items which turned out to be too difficult.

In order to examine the reliability of a test, we use statistical tools such as classical item analyses and IRT. As a basic classical item analysis, facility values and discrimination indices are calculated. Alderson et al (1995: 80-81) state that "an item's facility value is the percentage of students to answer it correctly." And a discrimination index can give us the information on "how well it [an item] distinguishes between students at different levels of ability." (81) If an item shows a higher discrimination index, it discriminates better.

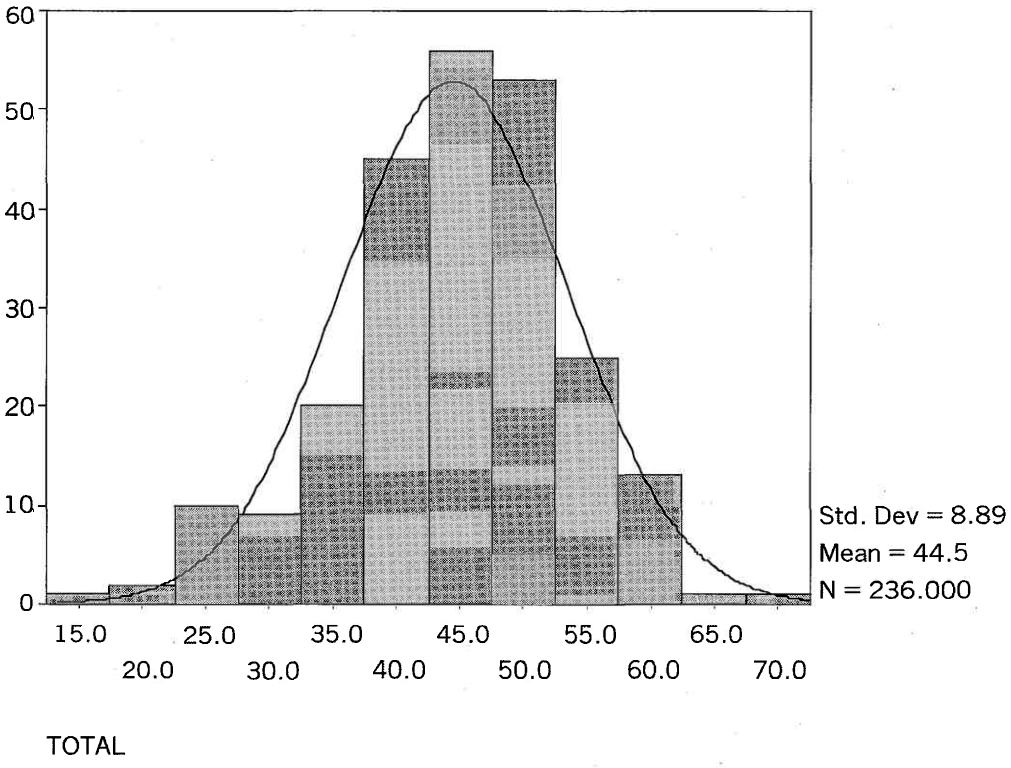
Since we use multiple choice items for our achievement test, another classical item analysis, 'analysis of distractors,' should be investigated to check how each distractor works. Some distractors attract more test takers than correct answers, and these items do not contribute to raise reliability of the test. Thus, we need to modify or delete these items.

Now we have three kinds of classical item analyses. Are these satisfactory enough? In order to investigate the test more thoroughly, I carried out a new item analysis using IRT. I will show the results of each analysis and also limitations of each analysis. I found out that we would need diversified analyses to make decisions which items should be modified or deleted. Furthermore, from the result of distractor analysis, I suggest that we can reduce the number of distractors.

3.2.3 The results of item analyses

Graph 1 below shows the histogram of the achievement test, from which we get the mean (average score) 44.5 (out of 75), which means 59.3 out of 100 and the standard deviation 8.89.

number of the students



Graph 1: Histogram of the achievement test

[1] Classical facility value and discrimination index

I used Excel and SPSS for this analysis. In order to calculate discrimination indices, we need to decide who belong to the top group, and who to the bottom group. I added 8.89 (standard deviation) to the mean 44.5 and chose the upper group of 38 students. Then, I also distracted 8.89 from the mean and got the lower group of 38 students. There are 68% of the students between the scores 35.51 and 53.29, so the number 38 shows the top 16% and the bottom 16% students, respectively, out of the total 236 students.

$44.5 + 8.89 = 53.29$

$44.5 - 8.89 = 35.51$

$236 \times 0.16 = 37.76$ (The top group and the bottom group have 38 students each.)

Table 1 shows facility values, which show average percentage of correct answers, of the two

groups, top and bottom, and discrimination indices, which are the differences between facility values of the top group and that of the bottom group, and also facility values of all the students for each item. F.V. stands for facility values, and D.I. for discrimination indices.

Alderson claimed in his lecture in 2005 that the interpretations depend on the types and purposes of your tests. Since this test is an achievement test, it is preferable for the students to get 60% in the facility value of the total. I set the acceptable range of facility values as higher than 45%. Although the facility value higher than 90% is considered too high and the item should be modified or deleted, I included these items which had high facility values. Because this is an achievement test, I prefer to include easy items to increase our students' confidence. Easy items should be the first ones on the test. If lower-level students can answer the questions with confidence, they can move on and challenge the following questions.

As for discrimination indices, I decided to give this achievement test the range 0.2 ^ 0.6 while in the case of a proficiency test, the range higher than 0.3 would be chosen. From the table 1, I can assume we have 17 problematic items out of 75.

ItemNo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
F.V.of T	0.95	0.9	0.95	0.9	0.74	0.87	0.53	0.92	0.71	0.95	0.79	0.95	0.76	0.87	0.68
F.V. of B	0.63	0.53	0.61	0.47	0.18	0.53	0.32	0.63	0.71	0.45	0.37	0.61	0.42	0.34	0.34
D.I.	0.32	0.37	0.34	0.42	0.55	0.34	0.21	0.29	0	0.5	0.42	0.34	0.34	0.52	0.34
F.V.of all	0.87	0.74	0.83	0.75	0.50	0.75	0.41	0.88	0.75	0.73	0.64	0.71	0.64	0.61	0.54
Item No	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
F.V .of T	0.70	0.47	0.55	0.68	0.76	0.42	0.32	0.9	0.97	0.74	0.92	0.82	0.74	0.82	0.89
F.V. of B	0.34	0.29	0.32	0.34	0.34	0.18	0.03	0.53	0.68	0.21	0.66	0.29	0.24	0.55	0.4
D.I.	0.45	0.18	0.24	0.34	0.42	0.24	0.29	0.37	0.29	0.53	0.26	0.53	0.5	0.26	0.47
F.V.of all	0.53	0.29	0.38	0.61	0.54	0.33	0.14	0.75	0.88	0.54	0.85	0.45	0.5	0.69	0.66

Item No	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
F.V.of T	0.79	0.84	0.95	0.68	0.76	0.68	0.5	0.68	0.9	0.4	0.97	0.84	0.66	0.92	0.55
F.V. of B	0.47	0.58	0.61	0.16	0.5	0.29	0.24	0.37	0.42	0.16	0.66	0.29	0.24	0.47	0.34
D.I.	0.32	0.26	0.34	0.53	0.26	0.4	0.26	0.32	0.47	0.24	0.32	0.55	0.42	0.45	0.21
F.V.of all	0.62	0.66	0.87	0.35	0.66	0.48	0.34	0.5	0.68	0.30	0.87	0.57	0.53	0.73	0.49
Item No	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
F.V.of T	0.55	0.68	0.84	0.95	0.95	0.74	0.63	0.58	0.68	0.82	0.76	0.68	0.74	0.53	0.82
F.V. of B	0.55	0.24	0.4	0.47	0.37	0.32	0.34	0.26	0.58	0.37	0.45	0.34	0.32	0.26	0.45
D.I.	0	0.45	0.45	0.47	0.58	0.42	0.29	0.32	0.11	0.45	0.32	0.34	0.42	0.26	0.37
F.V.of all	0.55	0.48	0.63	0.68	0.73	0.53	0.43	0.49	0.57	0.62	0.65	0.5	0.6	0.46	0.72
Item No	**61**	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**
F.V. of T	1	0.9	0.97	0.61	0.61	0.92	0.76	1	0.71	0.92	0.45	0.45	0.9	0.55	0.79
F.V. of B	0.79	0.42	0.76	0.29	0.24	0.24	0.29	0.58	0.37	0.45	0.26	0.16	0.34	0.37	0.24
D.I.	0.21	0.47	0.21	0.32	0.37	0.68	0.47	0.42	0.34	0.47	0.18	0.29	0.55	0.18	0.55
F.V.of all	0.94	0.81	0.92	0.53	0.41	0.68	0.37	0.87	0.48	0.72	0.33	0.27	0.53	0.37	0.56

Table 1: Facility values and Discrimination indices of the achievement test

The items which are problematic in facility values are items 7, 17, 18, 21, 22, 34, 37, 40, 52, 65, 67, 71, 72, and 74. The items which are problematic in discrimination indices are items 9, 17, 46, 54, 71, and 74. Here we can say we absolutely need to modify or delete items 17, 71, and 74 because these are problematic in both types of analyses.

Next we need to delete items 9, 46, and 54 since these have too low discrimination indices, 0, 0, and 0.11, respectively. Then we may modify or delete items 7, 18, 21, 22, 34, 37, 40, 52, 65, 67 and 72 since these have very low facility values.

Items 61 and 63 have very high facility values, 0.94, and 0.92, respectively, and do not discriminate well, but we may keep them to make the students feel confident since this is an achievement test.

[2] Classical distractor analysis

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	87.3	8.9	3.0	3.0	5.9	16.9	16.9	1.7	14.0	7.6	13.1	18.2	20.8	25.8	19.9
B	8.5	12.3	83.1	6.8	50.4	3.0	39.0	88.1	75.0	18.6	63.6	10.6	63.6	13.6	54.2
C	3.0	74.2	6.4	15.7	4.7	74.6	41.1	9.3	11.0	73.3	22.5	70.8	15.7	60.6	25.8
D	1.3	4.7	7.6	74.6	39.0	5.5	3.0	0.8							
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
A	53.4	17.4	37.7	3.0	0.4	7.2	16.1	5.1	3.8	12.3	2.5	10.6	50.0	26.3	66.1
B	25.0	28.8	27.5	31.8	5.9	32.6	13.6	75.4	1.7	23.7	85.2	44.9	16.1	4.2	20.8
C	21.6	53.8	34.7	61.0	39.0	24.2	58.5	8.9	87.7	10.2	7.2	25.4	12.7	0.4	1.3
D				4.2	54.2	35.6	11.9	10.6	6.8	53.8	5.1	18.6	20.3	69.1	11.9
	31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
A	28.0	65.7	3.4	7.2	65.7	47.5	22.0	49.6	10.2	30.1	1.3	19.1	21.2	8.9	7.6
B	62.3	11.9	86.9	34.7	6.4	7.6	18.6	23.3	68.2	15.7	11.0	57.2	53.4	4.7	49.2
C	6.4	8.1	1.7	16.5	21.2	16.9	25.4	24.6	9.3	14.4	86.9	11.9	5.9	14.0	19.1
D	3.0	14.4	8.1	41.5	6.8	26.7	33.5	2.5	11.9	39.0	0.8	11.9	19.5	72.5	24.2
	46	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
A	55.1	47.9	10.2	67.8	5.9	52.5	24.6	49.2	19.5	11.0	65.3	5.9	59.7	13.6	5.1
B	7.2	12.7	11.0	24.2	7.2	34.7	43.2	17.8	57.2	61.9	8.5	14.4	8.5	30.5	71.6
C	31.8	30.9	15.7	3.8	13.6	5.1	24.6	27.5	16.1	3.4	14.4	30.1	28.8	45.8	2.5
D	5.9	8.5	63.1	4.2	73.3	7.6	7.6	5.1	6.8	23.7	11.9	49.6	3.0	10.2	20.8

	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
A	1.7	80.5	2.1	52.5	18.6	67.8	4.7	2.5	9.7	2.1	14.0	19.1	24.2	33.1	12.3
B	1.7	10.2	4.7	25.0	11.4	15.3	37.3	6.8	11.4	12.3	26.7	32.6	16.9	18.6	55.9
C	3.0	2.1	1.3	2.1	40.7	3.4	8.9	87.3	30.9	72.0	33.1	21.6	5.5	37.3	17.4
D	93.6	7.2	91.9	20.3	28.8	13.6	49.2	3.0	47.9	13.6	25.4	26.7	53.4	11.0	14.0

Table 2: The percentage of correct answers and distractors

From Table 2, I chose 7 items which should be checked. The 7 items (17, 21, 22, 34, 40, 67, and 72) seem too difficult, because their correct answers attract fewer students than their strong distracters.

[3] IRT [item difficulty and unexpected responses]

Finally, I used IRT (item-response theory) to confirm the results of classical item analyses, and also to check the item difficulty and unexpected responses. Weir (2005: 26) states that item-response theory models have become increasingly popular measurement tools in the past 35 years. These models use responses to item on a test or survey questionnaire to simultaneously locate both the items and the respondents on the same latent continuum." Figure 1 indicates the distribution of test takers on the left and that of items on the right. By looking at the distribution, we can easily grasp the levels of the test takers and the items.

Logit	Number of students	<more>	l <rare>						
4	person		+ item						
			I						
	(1)		I						
			I						
3			+						
			I						
			I	22					
			I						
2			+						
	(1) .		I T						
	(1) .	T	I						
	(3) #		I	72					
	(1) .		I	17					
	(10) . ###		I	40					
	(1) .		I	21	71	37			
	(13) . ####	S	I	34					
1	(13) . ####		+	18	67	74			
	(18) ######		I	7	65				
	(16) . #####		I	27	52				
	(27) #########		I	47	59	69	36		
	(25) . ########	M	I	28	5	45	53	57	38
	(15) #####		I	16	25	43	51	64	73
	(18) ######		I	15	20	46	75		
	(16) . #####		I	42	54				
0	(13) . ####		+ M	14	19	31	55	58	
	(12) ####	S	I	11	13	48			
	(4) . #		I	30	32	56	35		
	(4) . #		I	29	49	66	39		
	(1) .		I	12	44	60	70		
	(7) . ##		I	10	2	4	6	50	
	(3) #	T	I	23	9				
	(1) .		I S						
−1	(3) #		+	62					
	(1) .		I						

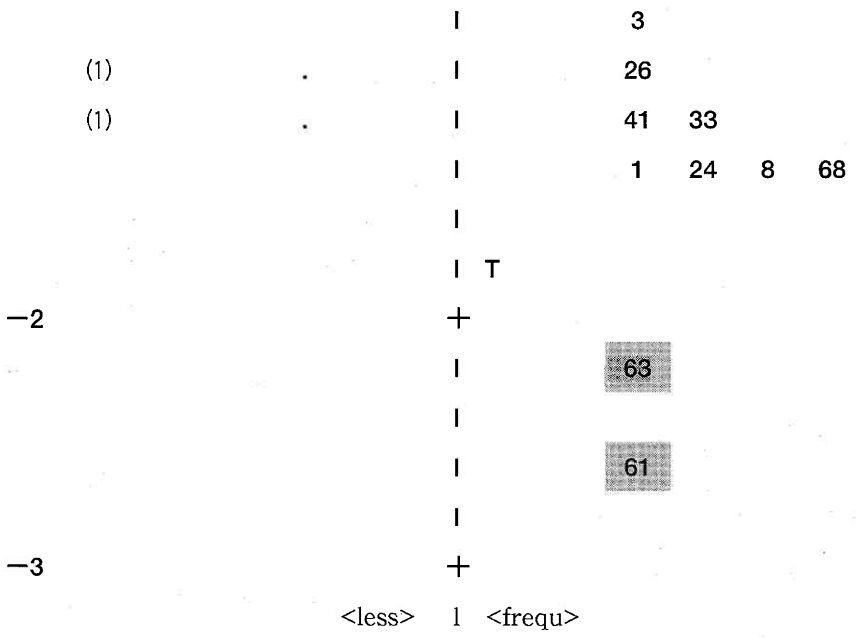


Figure 1: Persons map of items [Each '.' shows one person and each '#', 3 persons.]

Shown above is the figure of difficulty of items and the levels of test takers. I used the software Winsteps to get this map. Among the results by using the software Winsteps, this map is the most useful. We can see which items are too difficult or too easy, and can compare the difficulty of items and the levels of test takers.

Logit 0 shows the M (mean) of the items; that is, items 14, 19, 31, 55, and 58 are average items which have the same difficulty. The facility values are 0.61, 0.61, 0.62, 0.62, and 0.60, respectively. S shows one logit higher or lower, while T shows 2 logits higher or lower than M. The items which are higher than T are too difficult and the items lower than T are too easy. Here we can confirm the result of the classical item analyses. Those 'too easy or too difficult' items are supposed to be modified or deleted if this test is a proficiency test. However, in the case of achievement test, we can keep too easy items to make test takers feel confident as stated above.

The number of the test takers above logit 0 is 192 persons out of 236. Logit 0 shows there is 50% probability that the same level of test takers can choose correct answers of average items, This sounds the test is easy enough. But if we want a test easier as an achievement test, we need to check every item above logit 0.

Besides the difficulty of the items, Winsteps enables us to know most unexpected responses, which means that lower-level test takers get difficult items right while higher-level test takers get easy items wrong. Table 3 shows most unexpected responses in order in listening and reading.

most unexpected responses

Listening items	8	24	1	26	3	23	9	4	6	2	10	12	29	30
Reading items	61	63	68	33	41	72	50	70	60	39	49	66	31	58

Table 3 : Unexpected responses in order

Here we can get the information about some of the most unexpected responses we can take into consideration and decide whether we eliminate them or keep them.

Most difficult	T			S	M	S					T	Too easy	easiest
22								26	41	1		63	61
									33	24			
										8			
										68			

Table 4 : 12 problematic items based on IRT （gray）

Judging from the data of unexpected items and the item difficulty map, I chose 12 items to be checked. Item 22 was too difficult and an outlier which meant above T （logit 2), and items 61 and 63 were too easy and also outliers below T （logit -2). Item 72 turned out to be very difficult and also got more unexpected responses. Item 17, 40, 21 and 34 were rather difficult and also their correct answers did not draw the most test takers. Items 1, 24, 8, 68, 41, 33, 26, 3

were rather easy and got more unexpected responses.

Table 5 below shows the result of problematic items which were found out through 3 item analyses.

Listening items

Problematic items	1	7	8	9	17	18	21	22	24	26
Analysis [1] : 16		●		●	●	●	●	●		
Analysis [2] : 7					●		●	●		
Analysis [3] : 10	●		●					●	●	●

Reading items

Problematic items	33	34	37	40	41	46	52	54	61	63	65	67	68	71	72	74
Analysis [1] : 16			●	●		●	●	●			●	●		●	●	●
Analysis [2] : 7		●		●								●			●	
Analysis [3] : 10	●				●				●	●			●			

Table 5: Items to be checked after 3 item analyses

Analysis [1] , analyses of facility values and discrimination indices, has 16 problematic items, and analysis [2] , the distractor analysis, has 7 while analysis [3] has 10. In total we have 26 problematic items out of 75, which implies this test is weak.

Now how do we choose final problematic items? Do we have to choose items which are considered problematic at least in two of the three analyses? That is not the case, because apparently we have to modify or delete the items whose correct answers attracted fewer test takers than strong distractors, that is, items 17, 21, 22, 34, 40, 67, and 72. Only distractor analysis could find the problematic item 34. The items which did not succeed in discriminating well are item 9, 46, 54, 71 and 74, and these could not be found by the other two analyses, distractor analysis and IRT. Only IRT can find unexpected responses. Most of the items from the data of analysis [3] are the ones which show unexpected responses. So in this case we

can exclude these unexpected responses, and instead check extremely difficult items. We can keep the easiest and easier items because this test functions as an achievement test.

IRT also can show the difficulty of the items, which is very useful when we have a few items which have the same difficulty. In other words, if we have two problematic items which have the same difficulty, we can modify or delete one of them. Moreover, if we want to keep two or three items which have the same difficulty, it is easy to judge by using IRT.

From the analyses above, it is desirable to use classical analyses and IRT to ascertain which items should be modified or deleted. Each analysis can give different evaluations, so we need to be careful before we decide.

4. Discussion and a final decision

According to the results of the 3 item analyses, I conclude that none of the analyses is satisfactory on its own and we should use both classical item analysis and IRT in order to decide the items to be modified or deleted, because IRT cannot find all the items including the strong distracors which attract more students than correct answers. Items 17, 21, 22, 34, 40, 67, and 72 should be modified or deleted although IRT misses items 21, 34, and 67.

The 9 items which are so easy that all of the three distractors could attract only less than 10% of test takers might need a slight modification, or can be kept as they are if we put those items at the beginning of each section to make test takers feel confident.

Let us see the problematic items of each section. In our course, we focused on (1) photo section, (2) quick-response section and (4) grammar section, and spent less time for the other sections. In addition to the statistical data, we should take the content of the course into consideration.

(1) photo section : 3 problematic items (1, 7 and 8) out of 8

We can keep all of these items 1, 7 and 8, because we focused on the photo sections and

students can feel confident about what they have learned and they can get correct answers in this section fairly easily.

(2) response section : 3 items (9, 17, 18) out of 10

This section was done fairly well and is the only section which has 3 options while the other sections have 4 options. Item 17 should be modified because it has a strong distractor.

(3) conversation / announcement section : 4 items (21, 22, 24 and 26) out of 12

The most difficult item 22 should be deleted, and item 21 needs a slight modification. Item 24 can remain but needs to be moved at the beginning of the section because it is a rather easy item. Item 26 seems slightly easy, but it is acceptable enough for an achievement test.

(4) grammar section : 8 items (33, 34, 37, 40, 41, 46, 52 and 54) out of 30

We can keep the item 33, which is rather easy, to make the students feel better. Item 34 and 40 should be modified because of strong distractors. We keep item 37 without any change, because it has good distractors although it is rather difficult.

(5) reading comprehension section : 8 items (61, 63, 65, 67, 68, 71, 72, and 74) out of 15

As a whole, the items in the reading comprehension section are either too difficult or too easy. The result shows how difficult to develop good items. We can make items 61 and 63 more difficult or leave both of them as they are. We have to take out item 71 and 74 because of strong distracters.

Judging from the results and the discussion, I will modify or delete at least 17 items out of 26. In items 17, 21, 22, 34, 40, 67, and 72, the distractors performed too well. Items 9, 46, 54, 71 and 74 are problematic in discrimination indices, and cannot discriminate appropriately. Rather difficult items such as 7, 18, 37, 52 and 65 also should be simplified. The other items such as 1, 8, 24, 26, 33, 41, 61, 63 and 68 can be kept as they are, because these easier items can encourage lower-level students.

5. Conclusions and suggestions to make tests more useful

As I quoted in the introduction, Bachman & Palmer (1996) state that usefulness of tests is the most important. And Alderson, in his lecture on language testing in 2005, claims clearly that (construct) validity is the most important and reliability is useful to validate a test. In this paper, I focused on how we can increase reliability of items through two kinds of item analyses.

Through the item analyses, I came to two conclusions. First, we should use both classical item analyses and IRT. Second, we can reduce the number of distractors from three to two. I will justify these two suggestions in the following sections.

Classical item analyses and IRT are complementary, and neither of them is satisfactory enough to make a final decision about the items. Hughes (2003) emphasizes that "... both classical analysis and Rasch analysis [IRT] have contributions to make to the development of better tests."

To use one test for multiple purposes, we need to consider about the acceptable range of facility values and discrimination indices, reliability of items and the total, item difficulty and unexpected responses. As an achievement test, we accept rather high facility values, and at the same time as a proficiency test, we need to check discrimination indices and standard deviation so that we can judge that each item discriminates well and the test does not cluster.

In spite of the implementation of the pilot test, we still have 26 items, out of 75, which need to be deleted or modified. This fact shows that our test is rather weak. By using different kinds of item analyses, classical and IRT, we could find these 26 problematic items. After we decide which should be deleted or modified, we need to retest the modified items until we get acceptable evidences. This procedure increases the reliability.

The second conclusion is related to reliability and also practicality of tests. To make our test

more reliable and more valid, in addition to the use of two kinds of item analyses, we need to have clearer test specifications from the initial design process. Weir (2005: 14) states that a test should "always be constructed on an explicit specification, which addresses both the cognitive and linguistic abilities." However, these cannot be considered 'satisfactory.'

As Hughes (2003: 63) says, to write successful items "is extremely difficult" under the condition that we have limitations of time and staff. In the case of our test, out of 225 distractors, 88 distractors (39%) attracted less than 10% of test takers. This shows that it is highly difficult to make plausible distractors, so reducing the number of distractors is desirable in terms of other reasons as well; we have limited time and lack of staff members to make a test.

In his article titled *Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research,* Rodriguez (2005) empirically proved that to reduce the number of distractors has no influence on the reliability of the tests.

Rodriguez recommends using three options, which include one correct answer and two distractors, because:

1. Less time is needed to prepare two plausible distractors than three or four distractors.
2. More 3-option items can be administered per unit of time than 4- or 5-option items, potentially improving content coverage.
3. The inclusion of additional high-quality items per unit of time should improve test score reliability, providing additional validity-related evidence regarding consistency of scores and score meaningfulness and usability. (2005: 11)

Moreover, Shizuka et al. (2006) also reported that the reduction of distractors (three to two) gave no influence on the reliability of their entrance examinations. They state that "using three options instead of four did not significantly change the mean item facility or the mean item discrimination. Distractor analyses revealed that whether four or three options were

provided, the actual test-takers' responses spread, on the average, over about 2.6 options per item, that the mean number of functioning distractors was much lower than 2, and that reducing the least popular option had only a minimal effect on the performance of the remaining options. These results suggested that three-option items performed nearly as well as their four-option counterparts." (2006: 35)

With respect to practicality, reducing the number of distractors will decrease our workload which occurs when we develop test items. We always have trouble producing 3 plausible distractors, and also take time when we check them, because we usually check items four or five times. Moreover, we always get annoyed by third distractors which are usually strange or not attractive. Therefore, adopting two distractors instead of three, we can reduce mistakes and make the testing procedure (developing, implementing and evaluating) more effective and practical.

It is extremely difficult to make useful tests, but having clear purposes of tests, using two kinds of item analyses, and reducing the number of distractors will help us develop more reliable, more valid, and more useful tests.

References

Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L.F. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L.F. & Palmer, A.S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Brindley, G. (1986). *The Assessment of Second Language Proficiency: Issues and Approaches.* Adelaide: National Curriculum Resource Centre Adult Migrant Education Program Australia.

Hughes, A. (2003). *Testing for Language Teachers. Second Edition.* Cambridge: Cambridge University Press.

McNamara, T.F. (1996). *Measuring Second Language Performance.* London: Pearson Education.

Nunan, D. (1999). *Second Language Teaching & Learning.* Boston: Heinle & Heinle Publishers.

Rodriguez, M.C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice.* Summer 2005: pp.3-13

Shizuka, T, Takeuchi, O, Yashima, T, & Yoshizawa K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing,* 1 2006; vol.23: pp. 35-57.

Weir, C.J. (2005). *Language Testing and Validation.* New York: Palgrave Macmillan.