Analysis of Objective Test Items:
Towards a Revision of the Placement Test

SHIMIZU Yuko

要約：言語テストの開発では，目的，対象，内容等の設定に始まり，アイテム・バンクの構築を行い，予備テストとその結果分析を経て，より洗練されたテストへと改良する過程を繰り返していくのが原則である。立命館大学の経済・経営両学部の英語プログラムでは，一回生の4月に独自開発によるプレイスメント・テストを実施しているが，開発過程の中で実施すべき分析が行われないままで現在に至っている。しかし，信頼性や妥当性等のテストが備えるべき条件を満たした＜理想に近いテスト＞を実施していく上で，現行テストに関して，何らかの見直しが必須である。そこで，今回，テスト開発のひとつの段階を押し進めるために，1998年4月に実施したプレイスメント・テストの応答を用いて項目分析等の統計処理を行い，その結果をもとに部分的な改訂を試みた。本稿では，項目分析の結果とプレイスメント・テストの改訂内容の概要を示すと共に，今後の方向性について考察する。

1. Introduction

A few decades ago, Andrew D. Cohen (1980) presented in the first chapter of his book an overview of classroom testing in an intriguing way. The chapter was titled "The wh-questions of Classroom Testing," which was divided into several sections subheaded as *Why Test?, When to Test?, What to Test For?* and the like. Revisiting the chapter gave me an opportunity to think over what those subheadings implied when the placement test was administered to freshmen at the Colleges of Economics and Business Administration at Ritsumeikan University.

Some answers to the wh-questions are easily found. The placement test (the PT) is aimed to obtain information to place the new students at the appropriate level of our English program in accordance with their English abilities, which answers the questions of *Why Test?* and *What to Test For?*. Since the PT has to be conducted to the new students before the first class starts, the answer to *When to Test?* is inevitably decided.

The PT was originally developed by the faculty of College of Business Administration in the beginning of the 1990s and has been used ever since. We can tell what the test battery consists of by seeing the actual test book, which answers the question, *How to Test?*. Finally, Cohen subheaded *How to Evaluate the Test?* and the only information available to respond to this question is the students' raw scores on the PT.

This paper is directly related to the unanswered question: *How to Evaluate the Test?*. Using the test data of the PT in 1998, item analysis was conducted to examine the PT for the purpose of making test items more sound and fair to the test takers. It is not my intention to criticize the present test format developed by our colleagues, but to compensate for some necessary steps in the test development, which has long been neglected since the original version was constructed.

Prior to showing the result of the analysis, I overviewed some critical issues concerning the test development, all of which encouraged me to consider and apply the Cohen's wh-questions in the actual situation of our English program.

2. Test Development and the PT

2.1. Basics of Test Development

There are several steps to pursue in developing language tests. Carroll and Hall (1985) present the cycle of construction of developing and using tests. They divide the cycle into the following four phases: (1) design; (2) development; (3) operation; and (4) monitoring. The main features of these phases are summarized in Table 1.

To construct an ideal test, all those phases should be undergone in sequence. We must amass information over time that will tell us whether the test is doing its work with accuracy and effect. Guided by this information, we should decide what modifications to make to our original test and, if necessary, go through the whole cycle of construction and revision of the test again (Carroll and Hall, 1985, p.7). In our case, we have recently started to face a phase of monitoring. To be more precise, the PT had long stayed at the stage of operation, or on a pretesting stage without trialing review. Therefore, item analysis and other statistical operations of the PT are essential in the monitoring phase of rejecting, revising and adding items.

2.2. The Placement Test for the Colleges of Economics and Business Administration at Ritsumeikan University

Before discussing item analysis, we should know what a placement test is in general, and what the PT of our English program is like.

Brown (1996) divides language tests into two categories based on the types of decisions to make. That is, classroom-level decisions and program-level decisions. Classroom-level decisions are to assess how much of the materials and set of skills taught in a certain course has been learned by the learners. Achievement tests and diagnostic tests are used for that purpose. Program-level decisions, on the other hand, are to compare the performances of test takers to each other. Examples of this category include entrance examinations, proficiency tests and placement tests.

Regarding a placement test, which is a concern in this paper, several scholars in language testing (Harrison, 1983, p.4; Huges, 1989, p.14; Alderson, Clapham & Wall, 1995, p.292) share the same points in their definitions. In short, a placement test intends to make homogeneous groups in terms of general abilities or skills of the test takers. Although both placement and proficiency tests try out *fairly* general material, Brown (1996) points out that the former must be more specifically related to a given program, so that it efficiently separates the students into level groupings within that program (pp.11-12).

The PT started by the College of Business

Table 1 Test Construction Phases

Phase 1: Design	Description of testee(s)
	Specification of settings, needs
	Statement of test tasks, topics
Phase 2: Development	Construction of draft test
	Trials of test
	Analysis of trials and test revision
Phase 3: Operation	Introduction of test for practical use
	Making decisions on test information
Phase 4: Monitoring	Survey of test administration
	Establishment of test measurement characteristics
	Preparation of test revision schedule

(From Carroll and Hall, 1985, pp. 7-8)

Administration to conduct the class activities and teaching more effectively, particularly for English speaking teachers. The PT has now been expanded through the English program of the College of Economics in addition to the College of Business Administration. Freshmen of both colleges take the PT a few days after entering the university, which is usually during the first week in April. The total scores are used to place the students into five levels of groupings: Super Advanced, Advanced, Upper Intermediate, Intermediate and Basic.

The PT consists of two subtests – Listening Section and Vocabulary & Reading Section. Table 2 summarizes the test battery.

Analysis of the PT and TOEFL-ITP (Shimizu, 1998) indicated that there were significant differences among the students of each level determined by the PT and that a correlation between the PT and TOEFL-ITP was seen fairly high when we looked at the whole scores, not the scores of the subtests. This implied that the PT was a valid tool for placement purposes. However, a wide range of variance on each of the subtests within the same level was observed, which left possibilities of misplacement of some test takers.

Those test analyses should usually be conducted under the condition that the instrument used is valid and reliable. It is a matter of course in the area of measurement and evaluation that testing instrument be pretested and refined before it is used for actual tests. With regarding to the PT, however, no item analysis has been yet conducted. Therefore, it is

urgent for our language program to monitor the present testing tool and revise it if necessary, in order to place each test taker into a proper level.

3. Item Analysis

3.1. Improvement of Test Items

The test item is a basic unit of a test. In multiple-choice items seen in many standardized tests such as TOEFL and TOEIC, for example, each item is easily recognized since it is discrete. Brown (1996) defines it as the smallest unit that produces distinctive and meaningful information on a test or rating scale (p.49). A test consists of many items, some of which are often grouped as subtests (e.g. listening, reading, grammar). When we place our students in certain levels, we usually look at the total scores of the subtests that the students have achieved. Each test item, however, must be reliable and valid since appropriateness of each individual item of the test contributes to the test score. To examine the appropriateness of the test, item analysis, which is "the systematic evaluation of the effectiveness of the individual items on a test" (Brown, 1996, p.50) is often conducted. The analysis provides information that the test developer can use in contemplating which items must be rejected, whether revisions must be required, and what those revisions might be (Green, 1998, p.34).

3.2. What is Item Analysis?

Item facility (IF) and item discrimination (ID) are

Table 2 The Test Battery of the PT

Listening (17minutes)	40 items	65 points
I . Answering questions	15 items	
II . Understanding statements	15 items	
III . Comprehending dialogues	10 items	
Vocabulary and Reading (35 minutes)	35 items	35 points
IV . Identifying synonyms	20 items	
V . Multiple-choice cloze	6 items	
VI . Reading comprehension	9 items	

traditionally used to calculate each item of the objective test to see its appropriateness (Brown, 1996; Alderson et al, 1995). IF is a statistical index to examine the proportion of test takers who correctly respond to a given item. The formula for calculation of IF is:

$$IF = N_{correct}/ N_{total}$$

where $N_{correct}$ = number of test takers answering correctly

N_{total} = number of test takers

An IF value ranges from .00 to 1.00. The higher the value is, the easier the item is.

Item discrimination (ID), on the other hand, shows the degree to which item separates the test takers who performed well from those who performed poorly. Usually the upper and lower 27 % are used for calculating ID. First, the IFs for the upper and lower groups are calculated separately for each item. Then the IF for the lower group is subtracted from the IF for the upper group on each item. An ID value can vary from +1.00 to -1.00. The following is a formula of ID:

$$ID = IF_{upper} - IF_{lower}$$

where ID   = item discrimination for an individual item

IF_{upper} = item facility for the upper group on the whole test

IF_{lower} = item facility for the lower group on the whole test

3.3. Interpretation

Once we obtain those two values of IF and ID, we must interpret what those numbers mean. As for IF, very easy or difficult items inform us little of the varying levels of ability of the test takers. Therefore we eliminate those items from the test. Oller (1979) shows items falling somewhere between about .15 and .85 of IF values are usually preferred (p.247). According to Brown (1995), ideal items will have an average IF of .50 with the highest available ID (p.69). Alderson et al. (1995) indicate that if the examiners want a wide spread of scores from an exam, they will

select items which are as near to an IF of .50 as possible (p.81). In reality, however, those that fall in a range between .30 and .70 of IF values are said to be applicable (Brown, 1996, p.70).

Regarding ID, high positive ID values are desirable. Oller (1979) indicates that usually a value of .25 or .35 is set as a lower limit on acceptable IDs (p.252). More strict scholars say that ID of .40 and above with IF values of .20 to .80 is fairly ideal (Ikeda, 1978).

In our item analysis of the PT, we set our guidelines as follows;

Item Facility   .30～.70

Item Discrimination   .30 and above

4. Results of Item Analysis of the PT

Item analysis was conducted using the PT data of freshmen from the College of Business Administration (n=739). Descriptive statistics and correlation studies indicated that the data of Business Administration were more reliable than of Economics, since the latter had some outliers. (Shimizu, 1998). Then, those who scored 53 points and above out of 100 were assigned as High group (Hi: n=202), and those who scored 37 and below as Low group (Lo: n=195), to examine item discriminability.

Table 3 Descriptive Statistics

	mean	sd	mn	mx	n
whole	45.352	12.521	11	88	739
Hi	60.589	7.328	53	88	202
Lo	30.036	5.445	11	37	195

Table 3 shows mean scores of the PT, which consisted of the two subtests — sections of Listening and Reading and Vocabulary (see Table 2), followed by standard deviations (sd), minimum scores (mn), maximum scores (mx) and the numbers of samples of the whole students and subgroups of Hi and Lo.

4.1. Listening Section

The Listening Section is comprised of three parts. Results of item analysis of each part were shown separately in Tables 4, 5 and 6.

Part I (Answering questions): We can see that only five items out of fifteen met the guidelines we set: Items 3, 6, 9, 11 and 15. As high values of IF indicated, Items 1 and 4 were too easy and did not discriminate good and poor listeners of English on that type of test. Those easy items may be used as lead-in items for the test takers to get used to the test format, but should be excluded from the actual scoring. Item 13, on the other hand, turned out to be very difficult (IF=.085). Its negative value of ID (ID=-.020) indicated that more Lo students answered right than did Hi students. Item 13 produced the unfavorable kind of variance after all. So it had to be eliminated from the test battery.

Table 4 Listening: Part I

item #	Hi (%)	Lo (%)	ID	IF
1	98.020	92.308	.057	.966
2	36.139	13.918	.222	.237
3	67.822	22.105	.457	.437
4	95.545	76.410	.191	.890
5	50.000	36.082	.139	.402
6	89.109	54.404	.347	.723
7	52.475	22.798	.297	.342
8	26.238	24.607	.016	.260
9	77.723	40.104	.376	.606
10	35.821	16.754	.191	.289
11	65.347	28.497	.369	.461
12	43.781	23.316	.205	.300
13	9.406	11.399	-.020	.085
14	44.059	23.560	.205	.346
15	64.851	26.943	.379	.453

Hi: Proportion of students in Hi who got the item right

Lo: Proportion of students in Lo who got the item right

ID: Item Discrimination   IF: Item Facility

Part II (Understanding statements): In Part II, Items 16, 19, 23, 25, 27 and 29 were qualified as good items. Items 18 and 21 did not satisfy IF, but discriminated

the two groups well. Therefore these items were reasonably good but were subject to improvement.

Table 5 Listening: Part II

item #	Hi (%)	Lo (%)	ID	IF
16	76.733	21.762	.550	.473
17	48.020	21.466	.266	.296
18	90.594	55.208	.354	.769
19	75.743	37.824	.379	.592
20	40.299	10.825	.295	.228
21	89.500	54.255	.352	.770
22	44.059	19.577	.245	.311
23	57.921	15.789	.421	.365
24	42.079	30.851	.112	.357
25	58.911	21.649	.373	.418
26	37.811	13.542	.243	.218
27	64.851	28.351	.365	.432
28	56.716	28.947	.278	.399
29	65.174	14.211	.510	.364
30	39.109	21.762	.173	.285

Part III (Comprehending dialogues): Five items out of ten met our criteria: Items 31, 34, 35, 37 and 40. Even two of the no-good items (Items 32 and 33) satisfied ID (ID=.357 and .403 respectively) with a little too high IF (IF=.759 and .748 respectively). Therefore we did not have to omit them but to change some of the distracters to make the items more difficult.

Although Part III required higher comprehension of spoken English than the other two parts in the Listening Section, item analysis showed that the items of Part III discriminated the test takers reasonably well and that their difficulty levels were appropriate. One thing we must bear in mind here is that in actural tests dialogues were repeated twice only in this part,

Table 6 Listening : Part III

item #	Hi (%)	Lo (%)	ID	IF
31	59.406	19.689	.397	.357
32	91.584	55.897	.357	.759
33	91.584	51.309	.403	.748
34	82.673	32.979	.497	.583

35	66.337	35.079	.313	.521
36	49.505	23.684	.258	.339
37	58.416	21.762	.367	.402
38	35.644	19.895	.157	.250
39	44.554	19.271	.253	.269
40	77.228	39.583	.376	.591

while other parts were heard only once, which possibly influenced the results.

4.2. Reading & Vocabulary Section

The Reading & Vocabulary Section is made up of three parts — Part IV through Part VI. Results of item analysis of each part were shown separately in Tables 7, 8 and 9.

Part IV (Identifying synonyms): Part IV required the test takers to identify the meaning of the underlined

Table 7 Vocabulary: Part IV

item #	Hi (%)	Loa (%)	ID	IF
41	49.505	38.021	.115	.431
42	59.901	19.271	.406	.379
43	22.772	18.519	.043	.183
44	87.562	43.684	.439	.682
45	91.584	43.979	.476	.739
46	72.772	31.383	.414	.552
47	24.257	15.426	.088	.204
48	39.801	21.164	.186	.281
49	79.208	47.090	.321	.686
50	63.366	42.105	.213	.551
51	61.881	37.766	.241	.505
52	90.594	53.125	.375	.773
53	68.317	20.430	.479	.465
54	47.525	30.319	.172	.350
55	49.505	17.021	.325	.378
56	45.545	35.829	.097	.393
57	46.535	34.921	.116	.429
58	58.911	28.723	.302	.458
59	47.030	20.942	.261	.344
60	45.274	17.989	.273	.320

word in a sentence and to choose a similar word or phrase from the given choices. As Table 7 shows, seven items met our criterion: Items 42, 44, 46, 49, 53, 55 and 58. However, the rest showed very low ID values, although seven of them had good IF values (Items 41, 50, 51, 54, 56, 57, 59, and 60). The purpose of the PT is to place the students in the right class level. If the test does not discriminate better and not-so-good learners of English, we must speculate the testing instrument and find a more proper instrument.

Regarding some items in Part IV, context was not given enough in the stem sentence. If the underlined word was totally an unknown word with little context for a test taker, there was no way to infer the meaning from the context, which probably forced him/her only to take a wild guess. Consequently, reliability of the test became low.

Part V (Multiple-choice cloze): Part V was intended to test reading comprehension using a cloze passage. The passage was 160 words in length and six words were deleted from it. The test takers were to choose an appropriate word to complete each deletion out of three choices.

Item analysis would not be a suitable operation in this type of test, since each item in this part was not discrete and depended on each other in context. We still examined the result while taking it as a premise. Therefore, it should be concluded that discriminability of Part V was very low, although IF values were all preferable and that adopting some different tools to test reading comprehension was recommended.

Table 8 Reading : Part V

item #	Hi (%)	Lo (%)	ID	IF
61	72.277	34.896	.374	.562
62	26.238	24.084	.022	.280
63	39.801	26.203	.136	.305
64	68.812	35.079	.337	.518
65	50.495	42.553	.079	.463
66	81.683	31.937	.497	.573

Part VI (Reading comprehension): Part VI was a typical reading comprehension test in a sense, which consisted of two different passages. Passage A was a 55-word advertisement with 4 questions (Items from 66 to 70) and Passage B was a 118-word expository text with 5 questions (Items from 71 to 75). Results in Table 9 showed that Passage B provided better items than Passage A in terms of both ID and IF values. This does not mean that we should simply omit Passage A. We must reconsider a reading test as a whole, including the previous parts of IV and V.

Table 9 Reading and Vocabulary Part VI

item #	Hi (%)	Lo (%)	ID	IF
67	98.020	73.684	.243	.898
68	78.218	28.042	.502	.574
69	66.832	27.807	.390	.479
70	32.178	19.372	.128	.258
71	51.485	21.277	.302	.355
72	81.188	45.455	.357	.688
73	93.069	40.323	.527	.719
74	67.822	16.848	.510	.429
75	90.500	35.870	.546	.619

5. How the PT can be Modified

Item analysis was one step in the test development, which meant that results were used to 'test a test.' By examining the results, the following five points were proposed and a modified version of the present PT was constructed.

(1) Listening : Parts I and II

The items which met our guidelines of ID and IF still remained. Other items with low ID and IF values were replaced with new items.

(2) Listening: Part III

We chose five better items from the original version and added one new type of listening task with five questions based on one long conversation. The reason we included the long conversation was that it would hopefully bridge academic listening (e.g. lecture listening), which our students would eventually encounter in their course work.

(3) Vocabulary: Part IV

The original items did not seem to be fair or sound as a vocabulary test. As a temporary treatment, therefore, new but analogous items with more contexts were constructed. The new items are not directed to measure vocabulary power but something of reading comprehension by encouraging the test takers to guess at the meaning of unknown words from the context. Further study and research on a vocabulary test is indispensable. With the establishment of theoretical framework, some vocabulary tests will be adopted in the future version.

(4) Reading Comprehension: Parts V and VI

The length of each passage used in the original version seemed to be too short to be called reading comprehension tests. Also, more items were needed to obtain valid results. Therefore we decided to use a cloze test, which is said to be a good testing instrument to measure overall language ability. (Oller, 1979) For practicality, we constructed a multiple-choice cloze test with 25 deletions since it allowed us to gain sufficient items in an economical manner.

To keep some aspects of the original version, five comprehension questions based on a 400-word expository passage were added.

(5) Test Battery

The original PT had two subtests: Listening Section and Reading & Vocabulary Section. In the revised PT, we added Grammar Section. Many proficiency tests and standardized tests such as TOEFL, TOEIC and G-TELP include grammar or structure sections in their subtests. This implies that there is some construct validity which can not be measured by reading and listening tests.

Empirically speaking, discrete-point items of grammar tests make it possible to process plenty of items in a short time. Therefore they will contribute to improving test reliability and yield variance among test takers.

There is another reason to adopt a grammar test. Poor learners of English are often troubled with bottom-up processing. By the grammar section, we may diagnose students' weak points. The information will be utilized to inquire and evaluate the present curriculum, teaching methods and textbooks in the long run.

In summary, the test battery of the revised PT consists of three sections as shown in Table 10.

6. Suggestions for Test Development

In the cycle of developing and using tests, as discussed in the earlier part of this paper, we must conduct a pretest before practical use. No matter how well modified and improved the PT may be, we cannot tell how it will work without being tested on a target population. Even experienced language teachers and testers are often unable to agree about what an item is testing or how difficult the item is for a given group of students (Alderson et al., 1995, p.73). Therefore it is necessary to try out the test on a small group of students as pilot testing before the major trials.

Our situation, however, does not allow us to conduct pilot testing. Therefore we reused many of the OK items from the original PT to eliminate maximum risk of adopting totally new test items without pretesting. The data that will be obtained from the main trial with new students this spring will be used for a placement purpose as well as for analysis and test revision as a next step.

Finally, I would like to make several suggestions for the next procedure of test revision and development.

(1) Use of IRT: Presently we use classical item analysis, which has a certain limitation. The results of the analysis are only true for those who actually took the test, since the test takers' characteristics and the test characteristics cannot be separated. Therefore the results may not be applied to other samples. The item analysis does not provide fixed measure of a test's difficulty. To overcome such a drawback, we need to use Item Response Theory (IRT), which makes it possible to develop an item difficulty scale. Several computer programs are now available so that we can easily obtain a test-person ability and item difficulty.

(2) Establishing Item Bank: Another merit of IRT is, as Alderson et al. (1995) mentions, that test developers can store items in item banks. 'Good' items can be calibrated according to item difficulty and other information obtained through IRT, and stored in an

Table 10 The Test Battery of the revised PT

Listening (22 minutes)	40 items	40 points
I. Answering questions	15 items	
II. Understanding statements	15 items	
III-1 Comprehending short dialogues	5 items	
-2 Comprehending a long dialogue	5 items \	
Grammar (20 minutes)	40 items	40 points
Reading and Vocabulary (35 minutes)	40 items	40 points
I. Multiple-choice cloze passage (15 min.)	25 items	
II. Reading comprehension (10 min.)	5 items	
III. Vocabulary (10 min.)	10 items	

item bank to be called upon when needed. An item bank usually consists of a few hundred items and surely requires time to build up. However, the bank will prove of enormous value and will save us a great deal of time and trouble later.

(3) Back to Basics: This will put us back to the starting point, but we must reconsider test specifications, which are the blueprint to be followed by test and item writers (Alderson et al., 1995, p.9). After 'monitoring' the test, we go back to the first phase of 'design.' We may need to identify characteristics of our students in terms of their language needs, English ability and the like. We must think of these questions again: what language skills and elements should be tested, what sorts of tasks are required, how many items are needed for each section, and what test methods are to be used?

7. Conclusion

In this paper, I have analyzed students' responses on the PT to obtain data for revising the present test items. The data were gathered by using classical item analysis procedures, and then the original PT was partially modified. Now a revised PT, or a quasi-new PT was developed. This is only the beginning of the test development. When we consider the PT or any types of tests *within* the language program, we must not forget about the relationship with its curriculum. I will conclude the paper with showing three sets of issues that Brown and Hudson (1998) proposed. First, we must realize importance of positive washback and find a way and ways to produce the effect by matching testing and curriculum. Secondly, the test scores should be interpreted diagnostically and used to inform the students of their strengths and weaknesses. Finally, it is important to use multiple sources of information to think about in selecting assessment

strategies and in interpreting their results (pp.667-671). Those will give us some implications to evaluate our language program in a long term perspective.

REFERENCE

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

Brown, J. D. (1996). *Testing in Language Programs*. Prentice Hall Regents.

Brown, J. D. & Hudson, T. (1998). The Alternatives in Language Assessment. *TESOL Quarterly*, 32. 4. 653-675.

Carroll, B. J & Hall, P.J. (1985). *Make Your Own Language Tests*. Pergamon Press.

Cohen, A.D. (1980). *Testing Language Ability in the Classroom*. Newbury House Publisher.

Green, A. (1998). *Verbal Protocol Analysis in Language Testing Research: A Handbook*. Cambridge University Press.

Harrison, A. (1983). *A Language Testing Handbook*. Macmillan Press.

Heaton, J.B. (1975). *Writing English Language Tests*. Longman.

Huges, A. (1989). *Testing for Language Teachers*. Cambridge University Press.

Ikeda, H. (1978). *Tesuto de Nouryoku ga Wakaruka*. Nihonkeizai Shinbun-sha.

Oller, John W., Jr. (1979). *Language Tests at School*. Longman.

Shimizu, Y. (1998). *Pureisumento Tesuto no Kekkabunseki*. Presented at Gengo Kyouiku Shisutemu Kenkyuushitsu Kenkyuukai No1.

Shimizu, Y. (1998). Analysis of English Placement Tests in a Language Program. *The Ritsumeikan Economic Review*, Vol. XLVII, No.2 · 3 · 4, 359-374