ラッシュ分析法によるプレイスメントテストの一考察
A Rasch-based Analysis of a Placement Test

中村　優治 NAKAMURA, Yuji

● 慶應義塾大学
Keio University

Keywords プレイスメントテスト，ラッシュ分析法，妥当性，信頼性，実用性
placement test, Rasch analysis, validity, reliability, practicality

ABSTRACT

Placement testing is probably one of the most widespread uses of tests within institutions. Though some institutions still choose commercially-produced proficiency tests as placement tests, an institution's placement test should be connected with its curriculum, and therefore, ideally the tests should be developed by the test users to meet their needs. The present research describes a reliability and validity study of a placement test (four sub-components: grammar, vocabulary, passage reading, and a cloze test) and provides an opportunity to discuss the research methodology. The test data was analysed using a Rasch model statistical program(RUMM). For reliability examination, we calculated a reliability coefficient which is the equivalent of KR 20 (an estimate of internal consistency) used in classical test analysis. For validity investigation, content validity (asking whether teachers thought that test content represented program context) was examined along with face validity (student perception of the test by an informal interview). The location order was examined to obtain the construct of the item difficulty order. The item characteristic curves were examined to check the discriminating power of each item.

プレイスメントテストは最近多くの大学で行われている．その中には市販のテストをこの目的のために利用しているところもあるが，特定の大学のプレイスメントテストはその大学のカリキュラムと密接に結びついているはずなので，理想的には当該の学生のレベル，大学の目標にあった独自のプレイスメント作りが行われるべきである．本稿は，大学で実施した読解力プレイスメントテスト（4つの下部セクションから成り立っている：文法，語彙，読解，クローズテスト）の信頼性，及び妥当性について検証結果をのべ，今後の研究の方向性について考察を加えようとするものである．テストデータ分析はラッシュモデル法によるRUMMプログラムを使用し，IRT手法による分析を行うと同時に古典的テスト理論によるKR20を用いて信頼性の検証も行った．妥当性の検証においては内容妥当性及び表面妥当性を考慮にいれ，また，項目分析は項目難易度及び項目特性曲線の観点から検証を行った．

1. Introduction

The Faculty of Letters at Keio University primarily aims to improve students' reading ability to further enhance their college learning. For that purpose the development of a placement test is needed to accurately place the students into their appropriate proficiency levels for a better learning experience, and to offer enough activities to enhance students' multi-faceted English communication ability.

The purpose of the placement test of the faculty of letters is to measure their English reading ability and to collect information on their proficiency in order to make classes appropriate to their English reading ability." The immediate goals are as follows:

1) to make classes according to their English reading ability
2) to offer classes for those who need remedial instruction
3) to offer classes for those who already are at the required level that need advanced classes to continue their further study

Although there are some existing commercial tests such as TOEFL-ITP, TOEIC-IP, G-TELP, EIKEN(STEP), and CASEC, it was agreed among the faculty members that the content, the level and the purpose of those tests are not appropriate for the placement test of the literature department. Furthermore, the results of the admissions test cannot be used for any other purpose than the entrance examination selection. For these reasons, we have decided to develop our own placement test.

Westrick (2005, p.90) says:

More studies on the use of commercially-produced tests and in-house tests for placement purposes at other Japanese colleges and universities are needed. Creating an effective placement test involves developing test items related to a true curriculum with clear goals and objectives, piloting the tests items, analyzing the data, and revising the tests to ensure that the scores are reliable and sound placement decisions can be made. This requires hard work, but it must be done if fair and defensible placement decisions are to be made.

Furthermore, the following scholars take a similar stance about the placement test. Brown (1996) says that a placement test must be more specifically related to a given program. Hughes (2003) claims that placement tests should be developed by the users themselves so that they specifically meet their needs. And, Fulcher (1997, p.113) argues, "The goal of placement testing is to reduce to an absolute minimum the number of students who may face problems or even fail their academic degrees because of poor language ability or study skills."

2. Purpose of the Study

The purpose of the present study is to examine the pilot version of the placement test and whether we can proceed to the real version of the placement test in a similar format.

McNamara (2000, p. 83) states, "There are three basic critical dimensions of tests —validity, reliability, and feasibility, whose demands need to be balanced." McNamara (2000, pp.50-51) also mentions three aspects that can threaten test validity:1) test content, 2) test method and 3) test construct.

Taking these three facets of a test into consideration, the research question for this study is the following: Does the Pilot version of this particular placement test have enough validity, reliability and practicality to proceed to the real test? This question gives rise to the following presuppositions :

Presupposition 1: The test has enough validity.

Basically, the validity can be examined whether the results fit the model or not.The construct validity in the Rasch model is investigated through the examination of five steps: 1) Chisquare examination, 2) Fitsresidual examination, 3) Location examination, 4) Item Characteristic Curves, and 5) Targetting information. Among these, the item analysis using the Item Characteristic Curves (ICC) is the main focus of this present research because this can make a great contribution to a better improvement of the revised test. Along with the ICC, the information of distractors will be discussed as well.

The content validity and the face validity can also be discussed in a non-statistical way. The test has content validity if the questions reflect the course content or syllabus. Face validity indicates if the test takers think that the test is measuring their reading ability. In the discussion of content validity, the test construct and the test method are additionally discussed. The test construct will be discussed in terms of the construct of the difficulty order of the subsections. The test method discussion will focus on how the test was planned, administered and scored. The face validity will be investigated through questionnaire results.

Presupposition 2: The test has the acceptable reliability.

The reliability is investigated by the person separation index, which is equivalent to the cronbach alpha. The benchmark for the acceptable boundary is over 0.7. Pieces of information will also be given by misfitting items.

Presupposition 3: The test has enough practicality.

The practicality of the test can be examined mainly by the timing factor. The test method is discussed as well. The purpose of the pilot version of the placement test is to examine the above presuppositions under the research question.

3. Method

3.1. Subjects

809 freshman university students in the Faculty of Letters of Keio University

3.2. Materials/ Instruments

A placement test for measuring students' English reading ability as well as grammar and vocabulary knowledge. It has four components: grammar section (15 items), vocabulary section (10 items), reading section (3 long passages with five questions each), cloze section (10 items).

N.B. The reading section has three reading passages which are classified as beginning level, intermediate level, and advanced level in terms of the content, the topic, and the vocabulary level out of the teachers' teaching experience. The length of the passages are about 400-500 words. The cloze section was intended to measure their grasping ability from the context.

3.3. Procedures

Test Construction

The Construct of Reading Ability, in other words, what is the reading ability, was established mainly from the following five aspects:

1) the teachers' teaching experience
2) the reading section of other existing tests
3) linguistic theories
4) the needs of the Mita campus where students are required to read the major books and references for their study areas. In other words, the required reading ability at the Mita campus.
5) the text books that are actually used in their

study areas.

The materials were searched and selected in the following way.

1) The grammar items were chosen by taking into consideration almost all of the grammar items that were supposed to have been mastered at the high school level.
2) The reading passages were selected from the three viewpoints (humanities, social sciences and natural sciences), also taking into consideration the appropriate vocabulary level.

The final components of the present placement test were, as mentioned above, the four sections (grammar, vocabulary, reading and cloze).

Test Method,Test Format and Test Scoring

By taking into consideration the limitations of the nature of a placement test,i.e., administering the test at the busiest time of the academic year just after the entrance ceremony, and that scoring and informing the results had to be done very quickly, the test was a multiple-choice format rather than a constructed response format. The testing time was 60 minutes and the scoring was done using the optical mark reader in an objective way.

Test Analysis

The test data was analysed using the RUMM statistical program. The ChiSquare is investigated if there is a huge gap in the neighboring scores. The benchmark for the acceptable range for the FitResiduals is between -3 and 3. The location order is examined to obtain the construct of the item difficulty order. The item characteristic curves will be examined to check the discriminating power of each item. and, the distractor information will also be discussed. The benchmark for the person separation index of the test reliability is over 0.7.

Questionnaire for the face validity

"Do you think this placement test measures your reading ability effectively and appropriately?"

This questionnaire was used informally to ask about students' opinions of the test in order to check the face validity.

4. Results and Discussion

N.B. Legend: In the explanation below four types of abbreviations will be used. G stands for Grammar, V stands for Vocabulary, R for Reading and C for Cloze.

4.1. ChiSquare Probability Order

Table 1 ChiSquare Probability Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
25	V25	MC	0. 728	0. 075	1. 496	781. 97	4. 276	9	0. 892
2	G02	MC	-1. 191	0. 110	-0. 572	781. 97	4. 759	9	0. 854
23	V23	MC	0. 284	0. 078	0. 658	781. 00	6. 175	9	0. 722
8	G08	MC	-0. 448	0. 089	0. 661	781. 00	6. 283	9	0. 711
30	R30	MC	1. 353	0. 076	0. 042	781. 97	6. 548	9	0. 684
43	C43	MC	0. 802	0. 075	0. 976	780. 02	7. 183	9	0. 618
38	R38	MC	0. 209	0. 079	0. 215	779. 04	7. 787	9	0. 555
40	R40	MC	-0. 820	0. 098	-0. 874	778. 06	8. 037	9	0. 530
27	R27	MC	-1. 606	0. 127	-1. 153	781. 97	8. 169	9	0. 517
20	V20	MC	1. 483	0. 077	-0. 294	781. 00	8. 577	9	0. 477
31	R31	MC	0. 218	0. 079	1. 230	780. 02	8. 772	9	0. 458
48	C48	MC	0. 172	0. 080	0. 518	771. 21	9. 040	9	0. 433
33	R33	MC	0. 347	0. 077	-0. 186	781. 97	9. 051	9	0. 432
18	V18	MC	0. 405	0. 077	1. 989	778. 06	9. 579	9	0. 385
28	R28	MC	-3. 234	0. 253	-0. 868	781. 00	10. 653	9	0. 300
32	R32	MC	0. 533	0. 076	-0. 613	781. 97	10. 796	9	0. 289
5	G05	MC	0. 341	0. 077	0. 211	779. 04	11. 815	9	0. 223
36	R36	MC	-0. 034	0. 082	0. 430	774. 14	12. 346	9	0. 194
22	V22	MC	0. 748	0. 075	0. 978	776. 10	12. 365	9	0. 193
41	C41	MC	0. 484	0. 076	0. 962	780. 02	13. 116	9	0. 157
12	G12	MC	0. 057	0. 080	1. 072	780. 02	13. 304	9	0. 149
26	R26	MC	-1. 169	0. 109	-0. 631	781. 97	13. 323	9	0. 148
42	C42	MC	0. 902	0. 075	1. 848	778. 06	13. 526	9	0. 140
35	R35	MC	1. 499	0. 077	-0. 482	780. 02	14. 418	9	0. 108
45	C45	MC	-0. 526	0. 091	-0. 768	777. 08	14. 598	9	0. 102
15	G15	MC	-0. 024	0. 081	0. 012	780. 02	14. 703	9	0. 099
19	V19	MC	0. 723	0. 075	4. 998	780. 02	14. 843	9	0. 095
14	G14	MC	-1. 953	0. 145	-1. 503	781. 00	15. 122	9	0. 087
1	G01	MC	-0. 680	0. 094	-1. 467	781. 97	15. 424	9	0. 079
17	V17	MC	0. 978	0. 075	2. 545	780. 02	15. 440	9	0. 079
9	G09	MC	0. 652	0. 075	-0. 882	780. 02	15. 579	9	0. 076
21	V21	MC	0. 527	0. 076	-1. 753	781. 00	17. 097	9	0. 047
10	G10	MC	-2. 098	0. 153	-1. 534	781. 00	17. 766	9	0. 037
6	G06	MC	-0. 064	0. 082	-2. 258	781. 00	18. 563	9	0. 029
4	G04	MC	0. 672	0. 075	3. 148	781. 00	18. 594	9	0. 028
37	R37	MC	-0. 103	0. 083	-1. 914	779. 04	18. 754	9	0. 027
34	R34	MC	-1. 306	0. 114	-1. 958	781. 00	19. 300	9	0. 022
44	C44	MC	1. 116	0. 076	3. 518	771. 21	19. 936	9	0. 018
16	V16	MC	0. 975	0. 075	4. 733	780. 02	21. 973	9	0. 008
46	C46	MC	-1. 016	0. 104	-2. 359	775. 12	23. 194	9	0. 005
39	R39	MC	-1. 207	0. 110	-1. 976	778. 06	23. 730	9	0. 004
7	G07	MC	-0. 997	0. 103	-2. 294	781. 00	27. 909	9	0. 000
50	C50	MC	0. 103	0. 081	1. 497	764. 36	29. 045	9	0. 000
49	C49	MC	1. 975	0. 084	2. 342	764. 36	30. 129	9	0. 000
3	G03	MC	-0. 227	0. 085	-3. 647	780. 02	32. 560	9	0. 000
24	V24	MC	-0. 393	0. 088	2. 409	781. 97	35. 324	9	0. 000
13	G13	MC	-1. 311	0. 114	-0. 772	780. 02	37. 366	9	0. 000
47	C47	MC	2. 570	0. 095	2. 698	774. 14	45. 466	9	0. 000
29	R29	MC	-1. 702	0. 131	-3. 077	781. 97	45. 955	9	0. 000
11	G11	MC	1. 255	0. 076	5. 942	780. 02	61. 082	9	0. 000

N.B.

Seg : sequence Number

MC : Multiple choice

SE : standard Error

DF : Degree of Freedom

Chisg : Chisquare

Prob : Probability

This table shows that three items (C47, R29, G11) need to be examined because there is a gap in the neighboring items.

4.2. FitResidual Order

Table 2 FitResidual Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
3	G03	MC	-0.227	0.085	-3.647	780.02	32.560	9	0.000
29	R29	MC	-1.702	0.131	-3.077	781.97	45.955	9	0.000
46	C46	MC	-1.016	0.104	-2.359	775.12	23.194	9	0.005
7	G07	MC	-0.997	0.103	-2.294	781.00	27.909	9	0.000
6	G06	MC	-0.064	0.082	-2.258	781.00	18.563	9	0.029
39	R39	MC	-1.207	0.110	-1.976	778.06	23.730	9	0.004
34	R34	MC	-1.306	0.114	-1.958	781.00	19.300	9	0.022
37	R37	MC	-0.103	0.083	-1.914	779.04	18.754	9	0.027
21	V21	MC	0.527	0.076	-1.753	781.00	17.097	9	0.047
10	G10	MC	-2.098	0.153	-1.534	781.00	17.766	9	0.037
14	G14	MC	-1.953	0.145	-1.503	781.00	15.122	9	0.087
1	G01	MC	-0.680	0.094	-1.467	781.97	15.424	9	0.079
27	R27	MC	-1.606	0.127	-1.153	781.97	8.169	9	0.517
9	G09	MC	0.652	0.075	-0.882	780.02	15.579	9	0.076
40	R40	MC	-0.820	0.098	-0.874	778.06	8.037	9	0.530
28	R28	MC	-3.234	0.253	-0.868	781.00	10.653	9	0.300
13	G13	MC	-1.311	0.114	-0.772	780.02	37.366	9	0.000
45	C45	MC	-0.526	0.091	-0.768	777.08	14.598	9	0.102
26	R26	MC	-1.169	0.109	-0.631	781.97	13.323	9	0.148
32	R32	MC	0.533	0.076	-0.613	781.97	10.796	9	0.289
2	G02	MC	-1.191	0.110	-0.572	781.97	4.759	9	0.854
35	R35	MC	1.499	0.077	-0.482	780.02	14.418	9	0.108
20	V20	MC	1.483	0.077	-0.294	781.00	8.577	9	0.477
33	R33	MC	0.347	0.077	-0.186	781.97	9.051	9	0.432
15	G15	MC	-0.024	0.081	0.012	780.02	14.703	9	0.099
30	R30	MC	1.353	0.076	0.042	781.97	6.548	9	0.684
5	G05	MC	0.341	0.077	0.211	779.04	11.815	9	0.223
38	R38	MC	0.209	0.079	0.215	779.04	7.787	9	0.555
36	R36	MC	-0.034	0.082	0.430	774.14	12.346	9	0.194
48	C48	MC	0.172	0.080	0.518	771.21	9.040	9	0.433
23	V23	MC	0.284	0.078	0.658	781.00	6.175	9	0.722
8	G08	MC	-0.448	0.089	0.661	781.00	6.283	9	0.711
41	C41	MC	0.484	0.076	0.962	780.02	13.116	9	0.157
43	C43	MC	0.802	0.075	0.976	780.02	7.183	9	0.618
22	V22	MC	0.748	0.075	0.978	776.10	12.365	9	0.193
12	G12	MC	0.057	0.080	1.072	780.02	13.304	9	0.149
31	R31	MC	0.218	0.079	1.230	780.02	8.772	9	0.458
25	V25	MC	0.728	0.075	1.496	781.97	4.276	9	0.892
50	C50	MC	0.103	0.081	1.497	764.36	29.045	9	0.000
42	C42	MC	0.902	0.075	1.848	778.06	13.526	9	0.140
18	V18	MC	0.405	0.077	1.989	778.06	9.579	9	0.385
49	C49	MC	1.975	0.084	2.342	764.36	30.129	9	0.000
24	V24	MC	-0.393	0.088	2.409	781.97	35.324	9	0.000
17	V17	MC	0.978	0.075	2.545	780.02	15.440	9	0.079
47	C47	MC	2.570	0.095	2.698	774.14	45.466	9	0.000
4	G04	MC	0.672	0.075	3.148	781.00	18.594	9	0.028
44	C44	MC	1.116	0.076	3.518	771.21	19.936	9	0.018
16	V16	MC	0.975	0.075	4.733	780.02	21.973	9	0.008
19	V19	MC	0.723	0.075	4.998	780.02	14.843	9	0.095
11	G11	MC	1.255	0.076	5.942	780.02	61.082	9	0.000

According to the benchmark of the acceptable range (-3 to 3), among the three pointed out in the

Chisquare investigation, R29 is regarded as an overfitting(overdisriminating) item and G11 is considered as an underfitting (underdiscriminating) item.

So far, based on the Chisquare and FitResidual information, three items (R29 and G11 and C47) are pointed out as problematic. They need to be investigated further in the location order.

4.3. Location Order

Table 3 Location Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
28	R28	MC	-3.234	0.253	-0.868	781.00	10.653	9	0.300
10	G10	MC	-2.098	0.153	-1.534	781.00	17.766	9	0.037
14	G14	MC	-1.953	0.145	-1.503	781.00	15.122	9	0.087
29	R29	MC	-1.702	0.131	-3.077	781.97	45.955	9	0.000
27	R27	MC	-1.606	0.127	-1.153	781.97	8.169	9	0.517
13	G13	MC	-1.311	0.114	-0.772	780.02	37.366	9	0.000
34	R34	MC	-1.306	0.114	-1.958	781.00	19.300	9	0.022
39	R39	MC	-1.207	0.110	-1.976	778.06	23.730	9	0.004
2	G02	MC	-1.191	0.110	-0.572	781.97	4.759	9	0.854
26	R26	MC	-1.169	0.109	-0.631	781.97	13.323	9	0.148
46	C46	MC	-1.016	0.104	-2.359	775.12	23.194	9	0.005
7	G07	MC	-0.997	0.103	-2.294	781.00	27.909	9	0.000
40	R40	MC	-0.820	0.098	-0.874	778.06	8.037	9	0.530
1	G01	MC	-0.680	0.094	-1.467	781.97	15.424	9	0.079
45	C45	MC	-0.526	0.091	-0.768	777.08	14.598	9	0.102
8	G08	MC	-0.448	0.089	0.661	781.00	6.283	9	0.711
24	V24	MC	-0.393	0.088	2.409	781.97	35.324	9	0.000
3	G03	MC	-0.227	0.085	-3.647	780.02	32.560	9	0.000
37	R37	MC	-0.103	0.083	-1.914	779.04	18.754	9	0.027
6	G06	MC	-0.064	0.082	-2.258	781.00	18.563	9	0.029
36	R36	MC	-0.034	0.082	0.430	774.14	12.346	9	0.194
15	G15	MC	-0.024	0.081	0.012	780.02	14.703	9	0.099
12	G12	MC	0.057	0.080	1.072	780.02	13.304	9	0.149
50	C50	MC	0.103	0.081	1.497	764.36	29.045	9	0.000
48	C48	MC	0.172	0.080	0.518	771.21	9.040	9	0.433
38	R38	MC	0.209	0.079	0.215	779.04	7.787	9	0.555
31	R31	MC	0.218	0.079	1.230	780.02	8.772	9	0.458
23	V23	MC	0.284	0.078	0.658	781.00	6.175	9	0.722
5	G05	MC	0.341	0.077	0.211	779.04	11.815	9	0.223
33	R33	MC	0.347	0.077	-0.186	781.97	9.051	9	0.432
18	V18	MC	0.405	0.077	1.989	778.06	9.579	9	0.385
41	C41	MC	0.484	0.076	0.962	780.02	13.116	9	0.157
21	V21	MC	0.527	0.076	-1.753	781.00	17.097	9	0.047
32	R32	MC	0.533	0.076	-0.613	781.97	10.796	9	0.289
9	G09	MC	0.652	0.075	-0.882	780.02	15.579	9	0.076
4	G04	MC	0.672	0.075	3.148	781.00	18.594	9	0.028
19	V19	MC	0.723	0.075	4.998	780.02	14.843	9	0.095
25	V25	MC	0.728	0.075	1.496	781.97	4.276	9	0.892
22	V22	MC	0.748	0.075	0.978	776.10	12.365	9	0.193
43	C43	MC	0.802	0.075	0.976	780.02	7.183	9	0.618
42	C42	MC	0.902	0.075	1.848	778.06	13.526	9	0.140
16	V16	MC	0.975	0.075	4.733	780.02	21.973	9	0.008
17	V17	MC	0.978	0.075	2.545	780.02	15.440	9	0.079
44	C44	MC	1.116	0.076	3.518	771.21	19.936	9	0.018
11	G11	MC	1.255	0.076	5.942	780.02	61.082	9	0.000
30	R30	MC	1.353	0.076	0.042	781.97	6.548	9	0.684
20	V20	MC	1.483	0.077	-0.294	781.00	8.577	9	0.477
35	R35	MC	1.499	0.077	-0.482	780.02	14.418	9	0.108
49	C49	MC	1.975	0.084	2.342	764.36	30.129	9	0.000
47	C47	MC	2.570	0.095	2.698	774.14	45.466	9	0.000

R 29 is the closest to the easiest item (the fourth easiest) in the order, while G11 is the sixth most difficult one. C47 was pointed out in the ChiSquare order as being the most difficult one. Also, the location order shows that Reading items and Grammar items tend to be placed on the easier side of the continuum while Vocabulary items and Cloze items are on the relatively difficult side.

4.4. Item Characteristics Curves (ICC)

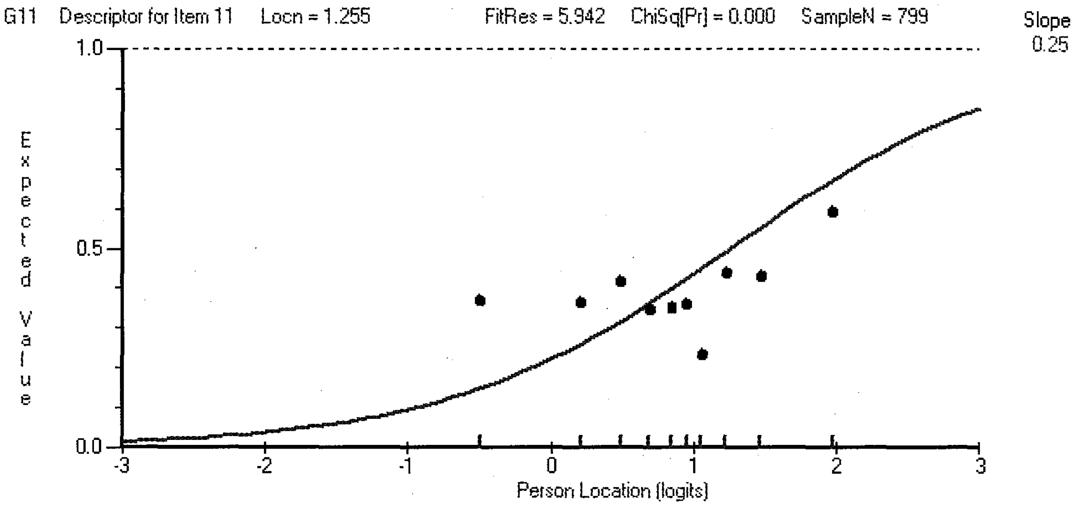

Figure 1 (G11 ICC)

This ICC of G11 shows us that the less able students perform better than anticipated. It also indicates that the more able students perform more poorly than anticipated. This further shows that this item does not discriminate the lower level students and intermediate level students probably because this item is a little too difficult (1.255 logit). This item probably would function better to differentiate the more able students at the top end. From the lower end to the mid group, there is no discrimination, even negative discrimination. From the mid to the top it has some discriminating power, but it is still not as much as anticipated.

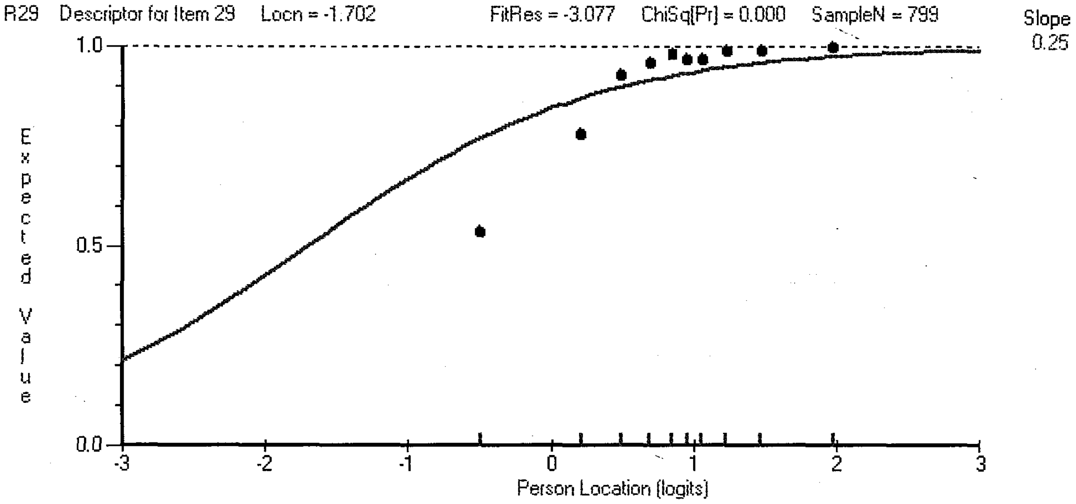

Figure 2 (R29 ICC)

This ICC of R29 shows that the item is problematic because it is overdiscriminating. We can tell the difference between the lower and the intermediate level students. However, it does not discriminate among the top level students. The top level students seem to have some advantage or bias about the topic. The less able students do not fit the model. The lower end is over discriminating, and the lower group is performing more poorly than anticipated.

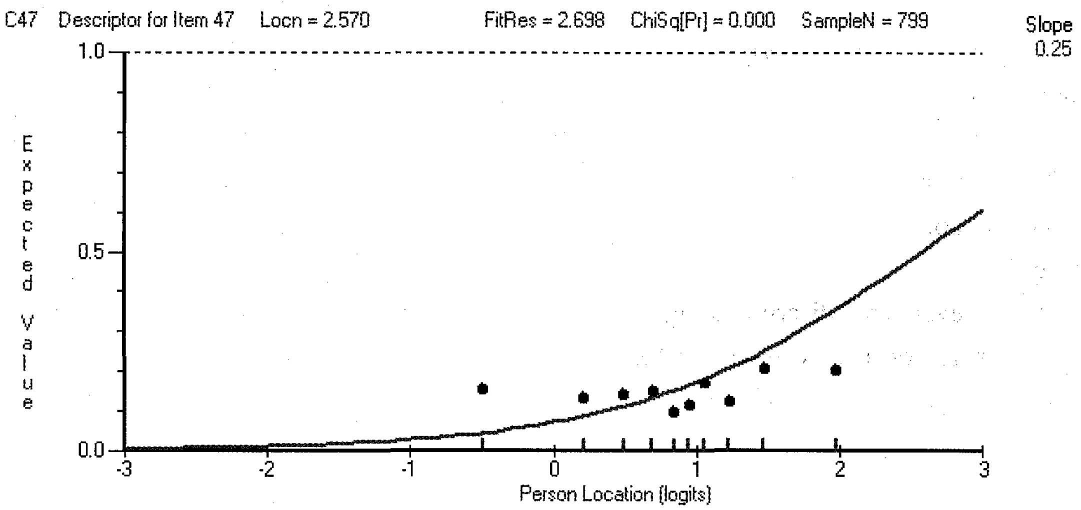

Figure 3 (C47 ICC)

This ICC of C47 indicates that the item has no discriminating power. All the groups get the item correct under the guessing level. This is probably the reason it was pointed out as problematic in the ChiSquare order.

So far, only three items have been pointed out as problematic. However, when we think about the percentage of these three problematic items, they are just three out of 50, only 6 % of the whole. Therefore, this figure is not overlly influential. Thus, we can say on the whole that the test is a reasonable measure.

N.B. Detailed information of each item ICC

Grammar Section

G1,2,5,6,8,9,10,12,14,15---Good

G3: The bottom end is poorer than anticipated by the mid end, while the top end is overdiscriminating.

G4: The bottom end is not working well (the item is good)

G7: The bottom finds this item harder. The bottom end is overdiscriminating.

G13: The lower end is overdiscriminating.

Vocabulary Section

V17,18,19,20,21,23,25---Good

V16: The lower end is not discriminating

V22: There are some flat parts, but on the whole this is still good.

V24: The lower end is almost not discriminating. The lower level students need a different sort of test.

Reading Section

R26,27,28,30,31,32,33,36,37,38,39,40---Good

R34: The lower end is a little overdiscriminating.

R35: The lower end is negative.

Cloze Section

C43,45,48---Good

C41: The mid groups are not discriminated.

C42: Good, but the lower end is not discriminating also.

C44: The lower end is negative and the mid group is negative, too.

C46: Too good.

C49: The lower level is not discriminating.

C50: This item is not a good measure for the neighboring groups.

4.5. Distractor Curve Information

Grammar Section

G1,2,4,5,6,7,8,12,13,14,15----Good.

G9: At the lower end, two distractors are more popular than the correct one.

G10: Good, but all the distractors are not distracting any of the students(It is an easy item, but easy items are necessary, too)

G11: See Figure 4 below. Strange. Up to 1.2, the key and the distractors are functioning in a confusing way. After 1.2 ability level, the key answer is functioning properly. Between 0.7 and 1.0, the students prefer option 2 to the key answer. The lower end and the upper end, in this case, got the item correct.

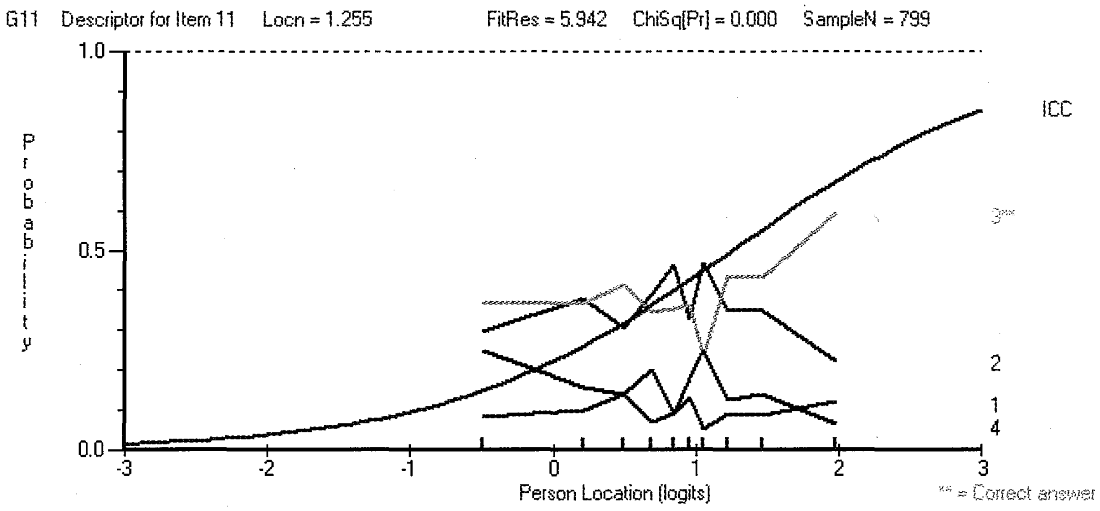

Figure 4 G11 Distractor Information Curve

Vocabulary Section

V16,18,19,21,22,23,----Good.

V17: Strange. Up to 0.8 ability level, the key answer does not work properly. The pattern is good and finally the key is on the right track.

V20: Strange. This is similar to V17. Up to 0.7 ability level, the distractor option 2 is more popular than the key.

V24: Easy item. But good.

V25: Up to 0.7 ability level, the distractor option 2 is more popular than the key. But this is a good item.

Reading Section

R26, 27,36---Good and easy.

R28,34,39,40----Easy or very easy

R32,33,37,38---Good.

R29: The key answer is functioning well and other distractors are less common than the key answer. This item is reasonable in terms of distractor functioning.

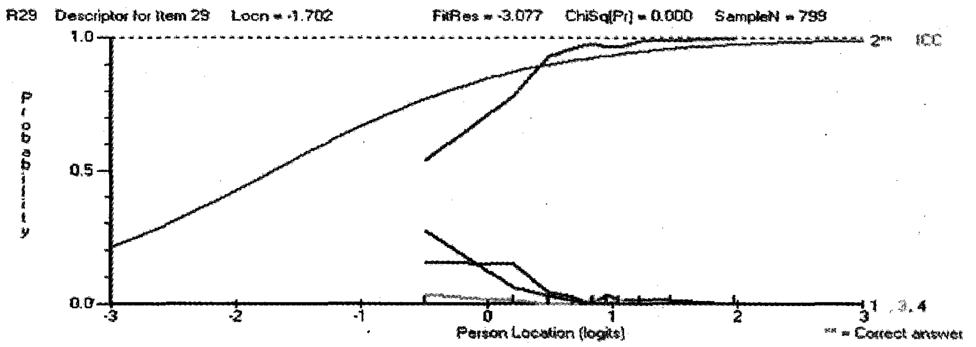

Figure 5 R29 Distractor Information Curve

R30: Good. Up to 1.2 ability level, the distractor is more popular. (a phenomenon of a Difficult item)

R31: Good. A phenonmenon of a difficult item.

R35: Good. A phenomenon of a difficult item. If we really want to measure an able person we need this item. In other words, also students after the 1.5 ability level can get this item correct.

Cloze Section

C41,42, 46,48,50----Good.

C44: Difficult

C46: Good and easy.

C43: Good. Up to the ability level 0.2, option 1 is the most popular. After 0.2 in all the groups option 4 is the most popular.

C47: See Figure 5 below. Strange. Difficult. The three distractors and the key answer do not function at any level. Even the key is chosen under the guessing level. It seems that there are two correct answers with option 3 as the most popular. All the students misunderstand the concept. The key answer is not discriminating.

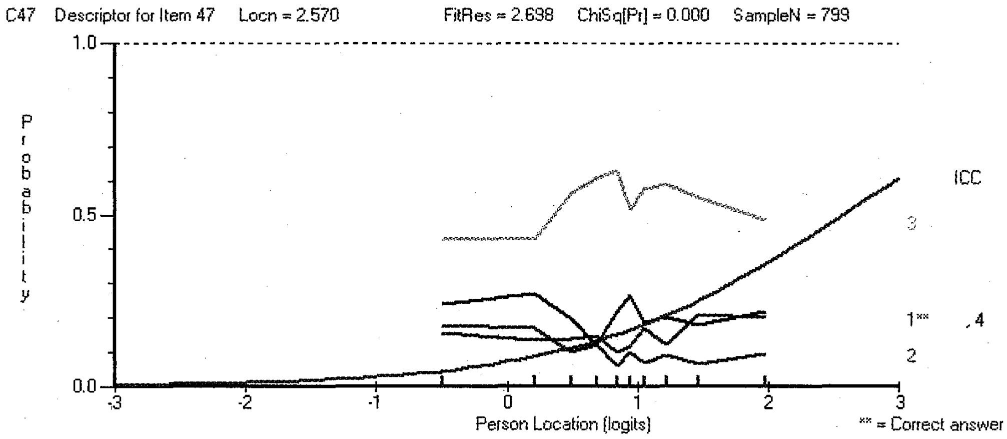

Figure 6 (C47 Distractor Information Curve)

C49: Strange. Good. Difficult. Up to 1.5 ability level, the distractors are more popular than the key. A phenomenon of a difficult item.

4.6. Information of Targetting

This figure suggest that as a whole the test is very good at measuring the students' English proficiency. For future improvement, we need more difficult items to match the more able students at the top end of this continuum.

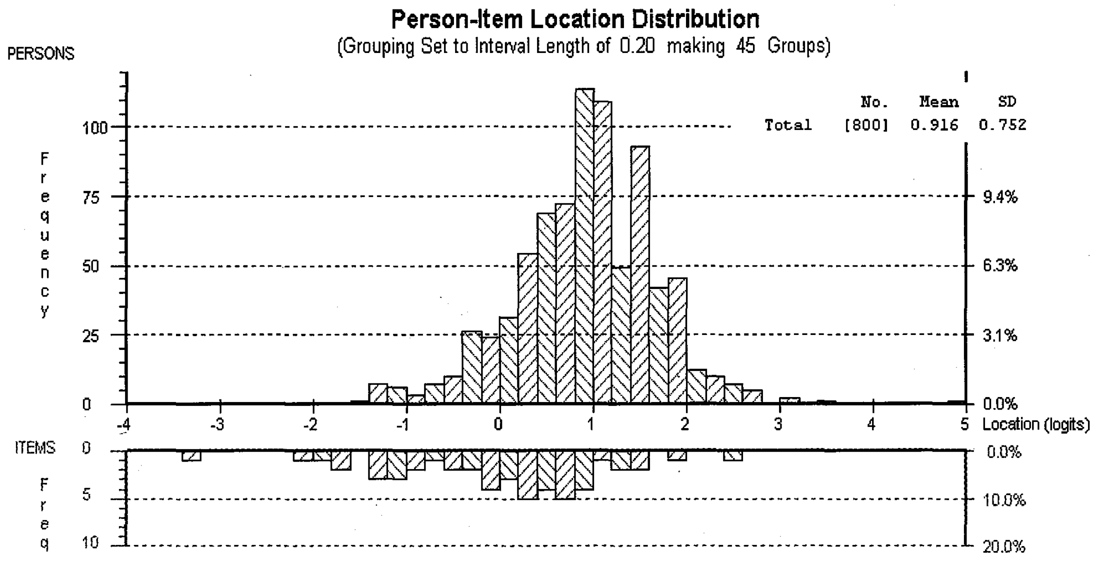

Figure 7 Information of Targetting

4.7. Examination of the reliability

The reliability was verified by the acceptable score of 0.78 in the person separation index in Table 4 below. This confirms the internal consistency of the items in this test.

Table 4 Reliability

RELIABILITY INDICES
Separation Index 0.786

4.8. Examination of the content and face validity

The content validity was verified through the discussion of the content of the test items. All the English teachers involved in this test development agreed to this test content. Furthermore, the construct validity was also investigated in the discussion of the test format and the content. The eventual test format was composed of the four subsections of English proficiency focusing on the reading ability.

The face validity was examined through the informal questionnaire and talks with the students by asking whether they had a feeling that they were taking a reading ability test. Most of the students agreed that the content of the test was that of a reading test.

4.9. Examination of the practicality

The practicality was supported by the test method and the whole process of the test administration. It took an hour to conduct the test and the results were analysed within the same day. The test was scored objectively.

4.10. Summary of the results and discussion

The Research Question for this study, "Does the Pilot version of a placement test have enough validity, reliability and practicality to proceed to the real test?" was supported by examining the three presuppositions as follows.

Presupposition 1, "The test has enough validity," was verified.

The finding of the three problematic items was a minor defect from the viewpoint of the whole test construct. In other words, almost all of the items fitted the model, illustrating the validity of the test.

In addition, the content validity of the test construction process and the face validity through the informal questionnaire results from the students support this statement.

Presupposition 2, "The test has the acceptable reliability," was verified by the person separation index. The reliability is investigated by the person separation index, which is equivalent to the Cronbach Alpha. The benchmark for the acceptable boundary is over 0.7.

Presupposition 3, "The test has enough practicality," was verified by information mainly about the timing factor of when the test was conducted and its successful implementation.

5. Conclusions and Implications

The Research Question for this study, "Does the Pilot version of a placement test have enough validity, reliability and practicality?" was partially supported with the examination of the three presuppositions. Also, the information obtained from the person-item relative position will help us divide the students into appropriate groups.

Considering McNamara's (2000, p.83) statement "The right balance will depend on the test context and test purpose," the present placement test should be acceptable judging from the statistical analysis and the test context as well as the test purpose.

For the future improvement, the predictive validity should be investigated as well.

Acknowledgement

The present author is grateful to Dr. David Andrich and Dr. Irene Styles for their invaluable comments and professional help.

References

Alderson, J.C. (2000). Assessing reading. New York: Cambridge University Press.

Andrich, D., Sheridan, B. & Luo, G. (2004). RUMM 2020: *Rasch Unidimensional Measurement Models..* Perth, Western Australia: RUMM Laboratory,

Bachman, L.F. (1999). *Fundamental considerations in language testing. Oxford*: Oxford University Press.

Brown, J. D. (2005). *Testing in Language Programs:* A Comprehensive Guide to English Language Assessment. New Edition. New York: McGraw-Hill.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing 14(2)*, 113-138

Grabe,W. (2000). *Reading research and its implications for reading assessment.* In A. Kunnan (Ed.), Fairness and validation in language assessment (pp.226-62). Cambridge: Cambridge University Press.

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

Linacre, M. (2004). *WINSTEPS Rasch Measurement computer program* (Version 3.51). Chicago, Winsteps. com.

McNamara, T.(2000). *Language Testing.* Oxford: Oxford University Press.

Westrick, P. (2005). Score Reliability and Placement Testing. *JALT Journal, 27(1)*, 71-92.