2016-2017 *ENGLISH LANGUAGE PLACEMENT TEST* (ELIPT) REVISION PROJECT

JAMES DEAN BROWN, HUY PHUNG, WEI-LI HSU, JONATHAN TRACE, KENTON HARSCH, & M. PRISCILLA FAUCETTE

University of Hawai'i at Mānoa

INTRODUCTION

Background

The main purpose of the research project was to analyze and revise the *English Language Institute Placement Test* (ELIPT) at the University of Hawai'i at Mānoa (UHM). All international students admitted to UHM are required to take the ELIPT before they register for courses at the beginning of their first semester of study (unless they meet the university's criteria for automatic exemption from the ELI). These students had previously reported their scores on standardized English proficiency tests (like the TOEFL or IELTS) as part of their application for admission to UHM. However, for placement purposes the ELI needs more detailed evidence of the students' language abilities in order to determine how the ELI could best meet their needs for support in English for academic purposes.

The English Language Institute

The English Language Institute (ELI) is housed in the Department of Second Language Studies (SLS) at the University of Hawai'i at Mānoa. Its primary goal is to provide academic English instruction for matriculated students who do not speak English as their native language—primarily international and immigrant students. Apart from providing instruction, the ELI also facilitates classroom research and observations of courses for department faculty and graduate students. Additionally, the ELI regularly serves as an advocate for students who have English as a second language, a group that is easily ignored and marginalized. Finally, the ELI provides consulting and expertise on matters related to second-language students to other offices and programs on campus.

The ELI staff includes two administrators, who jointly devote the equivalent of one full-time position to running the ELI. Previously, both administrators worked nearly full-time in the ELI. However, since the inception of the BA program in SLS, these two administrators split their time

Second Language Studies, 36(2), Spring 2018, pp. 1-25.

between the two programs. The ELI's instructional staff is made up entirely of graduate assistants, all of whom are MA or PhD students in SLS. The program also employs one full-time clerical staff member.

The ELI offers courses in three academic domains: listening/speaking, reading, and writing. Although these courses appear to separate the skills, in reality, each class integrates the four skills but focuses instruction on improvement in the designated skill area(s). These courses are designed to enhance students' awareness of academic discourse and expectations, and help students develop academic English abilities and strategies so that they can sufficiently participate in the academic environments of their various fields of study. In each domain, courses are offered in each of two levels: intermediate and advanced. Students who place into intermediate courses also take subsequent courses at the advanced level to fulfill the university's ELI requirements. Placement decisions are primarily based on the ELI Placement Test (ELIPT), with supplementary information (other academic test scores, students' prior language experiences, and educational backgrounds) used in borderline cases.

The English Language Institute Placement Test

The ELIPT is designed to measure students' academic English ability, and involves five separate sub-tests, including one writing placement test, two listening tests, and two reading tests. The complete test takes approximately four hours to complete, as follows:

- Writing Placement Test: 45 minutes
- Listening Test 1: Dictation: 10 minutes
- Listening Test 2: Academic Listening: 60 minutes
- 15-minute break
- Reading Test 1: Gap-Filling: 25 minutes
- Reading Test 2: Reading Comprehension: 55 minutes

In the writing domain, examinees select one of two given topics and write an argumentative essay in 45 minutes. They are given blank paper to make outlines and draft their essays but are not allowed to use dictionaries or other reference materials. Each essay is read by three different raters, who score the test according to a 4-category rubric. The average scores from three raters are used to decide students' placements. Students are placed into an intermediate writing course (ELI 73), an advanced course (ELI 83 for graduate students, or ESL 100 for undergraduates), or

are exempted from ELI writing requirements. Supplementary information (other academic test scores, students' prior language experiences, and educational backgrounds) is used to decide borderline cases.

For the listening/speaking domain, there are two sub-tests: a dictation and a multiple-choice academic listening test. In the dictation test, examinees listen to a 50-word recording three times, first at normal speed, then slower and with pauses, and last again at normal speed. They are asked to write every single word they hear from the recording. Spelling and punctuation is not graded, but grammatical knowledge is graded. The dictation is scored by one assessor. Each accurate word is counted as one point. In the academic listening test, examinees listen to three short and two long lectures and answer multiple-choice questions on a machine-readable answer sheet. There are 35 comprehension questions for this test. Based on these two tests, students are placed into an intermediate listening/speaking course (ELI 70), an advanced course (ELI 80), or are exempted from ELI listening/speaking requirements. Again, supplementary information is used to decide borderline cases.

For the reading domain, there are also two sub-tests: a 25-item gap-filling test and a 50-item multiple-choice reading comprehension test. In the gap-filling test, examinees have a total of 25 minutes to read a passage and use the information to fill in blanks on a summary and a chart. The question types are mainly designed to test the students' ability to identify, reorganize and summarize the key information from the passage. In the reading compression test, there are two sections: vocabulary and academic reading. In the vocabulary section, examinees are given a word or phrase and choose the option which has the closest meaning. In the academic reading section, examinees read six short passages and answer multiple-choice questions after each passage. All responses for the multiple-choice questions are recorded on machine-scored answer sheets. Based on these two tests, students are placed into an intermediate reading course (ELI 72), an advanced course (ELI 82), or are exempted from ELI reading requirements.

METHOD

Participants

Examinees in 2016. The examinees on the roster for the *2016 administrations* of the ELIPT consisted of 86 (39.1%) males and 133 (60.5%) females with one missing data point (0.4%) from

a variety of different nationality and language backgrounds. In terms of academic status, 157 (71.4%) were undergraduates and 63 (28.6%) were graduate students. Most of these students took the TOEFL Internet Based Test (IBT), Institutional TOEFL paper and pencil (P&P) test, or IELTS. Descriptive statistics for the scores of students who took these tests are shown on the left side of Table 1.

Table 1
Descriptive Statistics for Most of the 2016 and 2017 Examinees on the TOEFL Internet Based Test (IBT), Institutional TOEFL paper and pencil (P&P) test, or IELTS.

	2016			2017		
Statistic	TOEFL IBT	TOEFL P&P	IELTS	TOEFL IBT	TOEFL P&P	IELTS
N	75	17	19	90	15	23
M	90.31	518.59	6.00	83.61	540.27	6.17
Median	83.00	523.00	6.00	84.00	537.00	6.00
Mode	87.00	533.00	6.00	83.00	533.00	6.00
SD	54.13	25.93	0.46	9.37	20.57	0.32
Max	550	553	7	99	593	7
Min	61	430	5	61	500	6
Range	490	124	3	39	94	2

Examinees in 2017. The examinees on the roster for the *2017 administrations* of the ELIPT included 69 (50.0%) males and 69 (50.0%) females from a similar variety of nationalities and languages. In terms of academic status, 123 (69.5%) were undergraduates and 54 (30.5%) were graduate students. Most of these students took the TOEFL Internet Based Test (IBT), Institutional TOEFL paper and pencil (P&P) test, or IELTS. Descriptive statistics for the scores of students who took these tests are shown on the right side of Table 1.

ANALYSES AND REVISION PROCESS ORIGINAL 2016 ELIPT

The ELIPT revision process began in Fall 2016 when JD Brown took on the responsibility of Executive Director of ESL Programs. The first stage of the revision project involved organizing focus groups for Listening (ALT), Reading (RCT), Gap-Fill (GF), and Writing (WTG) tests. All ELI instructors and administrators were free to sign up for these focus groups. Rasch person/item maps and item fit statistics for each of the three tests were shown, explained, and discussed in

these meetings. The three focus groups were finished on October 10, 2016. The recommendations made by the focus groups for each of these tests will be discussed in turn.

2016 Academic Listening Test

For the ***Academic Listening Test*** (ALT) in 2016, Figure 1 shows the Rasch person/item map.

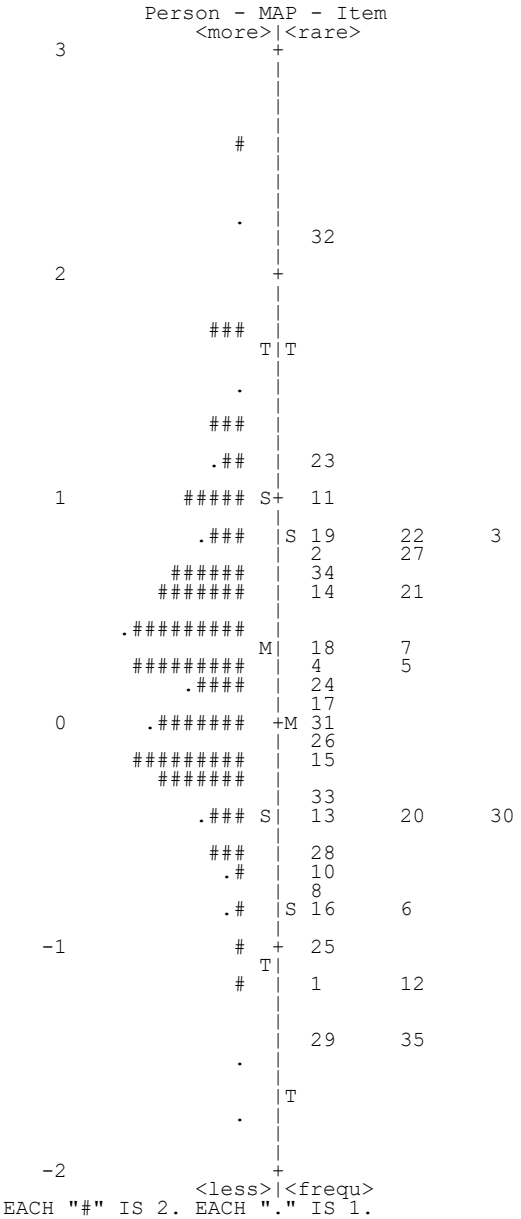

Figure 1. Person/Item Map for the 2016 ALT

For the ALT, the following problems were identified:

1. More difficult items were needed, especially at the upper level

2. Passage 1 should be replaced with a more difficult one. The new passage should be about science, such as nutrition, cancer, and medicine, including 400-500 words.
3. Passage 2 needed to be recorded in a more naturalistic way, with more false starts and pauses. But the original passage was retained.
4. Female voices should be included.
5. One item needed to be discarded, which led to the proposal that the new passage should have six items, instead of five.
6. The two new recordings would include ideas from the notions of World Englishes, so that the speakers' proficiency levels would be high but with detectable accents.

Revision decisions for the organization of the ALT version were as follows:

1. Dropping the first passage in the previous version of the test.
2. Moving the second and the third passages in the previous version so they became the first and the second passages in the revised version.
3. Adding the third passage in the current version.
4. Dropping Item 32 in the previous version from the current version.

In more detail, these revisions included:

1. Deleting the original Short Lecture 1 and making the original Short Lectures 2 and 3 the new Short Lectures 1 and 2. The original Short Lecture 1 had five items.
2. Re-recording the new Short Lecture 1 (about marketing, branding, and advertising) with a male local speaking Hawaiian Standard English.
3. Recording the completely new Short Lecture 3 (about modern biotechnology) with a high proficiency female Filipino accent. For Short Lecture 3, there were seven items. The recordings in numbers 2 & 3 were completed on November 14, 2016.
4. Deleting the original Item 32 from Long Lecture 2 (about alien civilization); based on item analyses, it was found to be too difficult and was therefore deleted from the revised ALT—leaving eight items for Long Lecture 2.
5. Table 2 shows the original ALT item numbers (on the right) and the numbers they became in the revised ALT (to the left).

Table 2

New and Original ALT Item Numbers

Original ALT Item #	New ALT Item #
6	1
7	2
8	3
9	4
10	5
11	6
12	7
13	8
14	9
15	10
16	11
17	12
New item	13
New item	14
New item	15
New item	16
New item	17
New item	18
New item	19
18	20
19	21
20	22
21	23
22	24
23	25
24	26
25	27
26	28
27	29
28	30
29	31
30	32
31	33
33	34
34	35
35	36

2016 Reading Comprehension Test

For the ***Reading Comprehension Test*** in 2016, Figures 2 and 3 show separate Rasch person/item maps for the reading (k = 25) and vocabulary (k = 25) items, respectively.

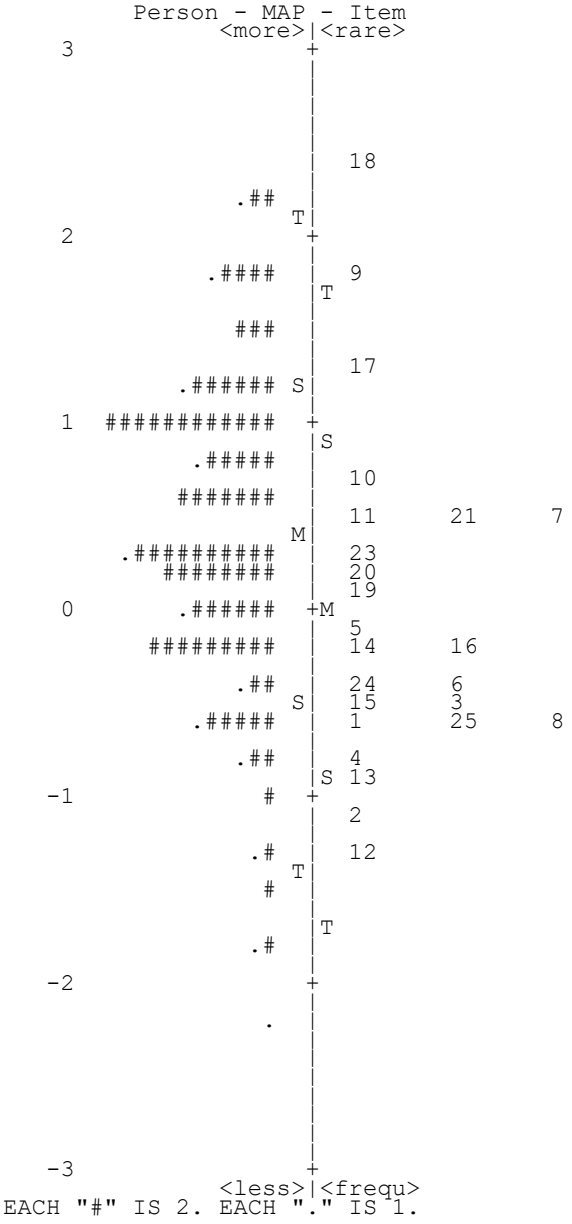

Figure 2. Person/Item Map for the 2016 RCT Reading

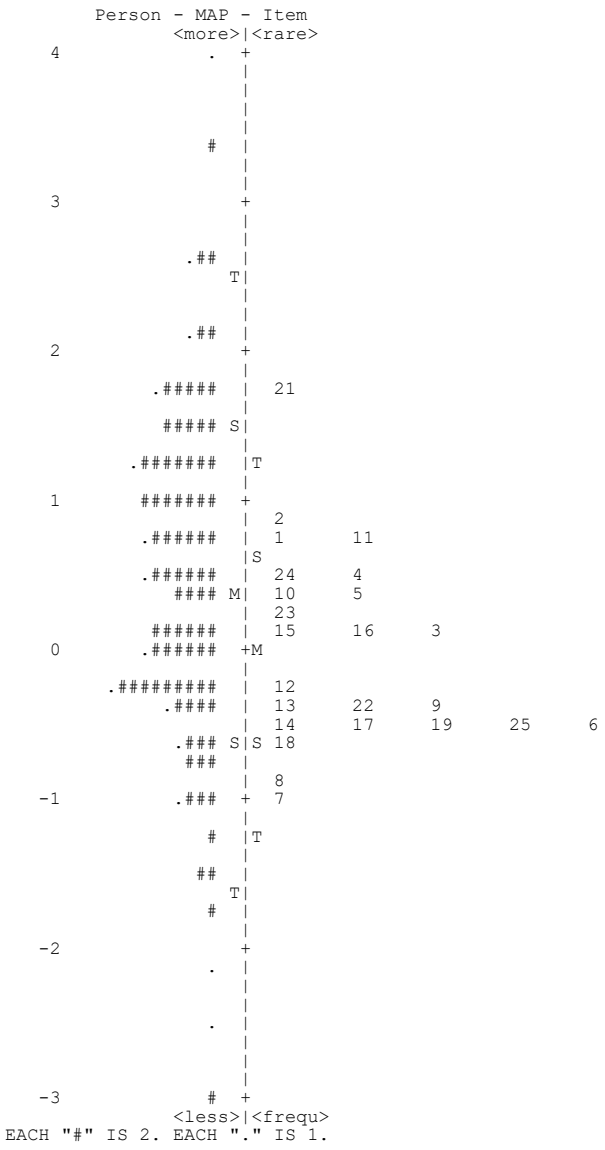

Figure 3. Person/Item Map for the 2016 RCT Vocabulary

For the ***2016 Reading Comprehension Test***, the following problems were identified:

1. The existence of more easy items than difficult items (based on Rasch analysis).
2. One item (new: Item 43) needed to be revised.
3. Vocabulary items need to be reordered and some easier items needed to be replaced with more difficult ones.

The resulting RCT revisions were as follows:

1. Moving 25 vocabulary items to the beginning of the test and the comprehension items to the end. Among the vocabulary items, the items were re-grouped based on the prompt

types. To be more specific, the decontextualized items were placed at the beginning, followed by the contextualized, synonym items, and the contextualized fill-in-blank items. Table 3 shows the original RCT item numbers (on the right) and the numbers they became in the revised RCT (to the left).

2. Revising Item 18 and moving it to Item 43 after re-ordering the items.

Table 3
New and Original RCT Item Numbers

New RCT Item Numbers	Original RCT Item Numbers
1	26
2	27
3	28
4	29
5	30
6	31
7	32
8	42
9	43
10	49
11	50
12	33
13	35
14	36
15	37
16	38
17	39
18	40
19	44
20	47
21	34
22	41
23	45
24	46
25	48
26	1
27	2
28	3
29	4
30	5
31	6
32	7
33	8
34	9
35	10
36	11
37	12
38	13
39	14
40	15
41	16
42	17

43	New item
44	19
45	20
46	21
47	22
48	23
49	24
50	25

Gap-Fill Test

For the ***Gap-Fill Test*** in 2016, the Rasch person/item is shown in Figure 4.

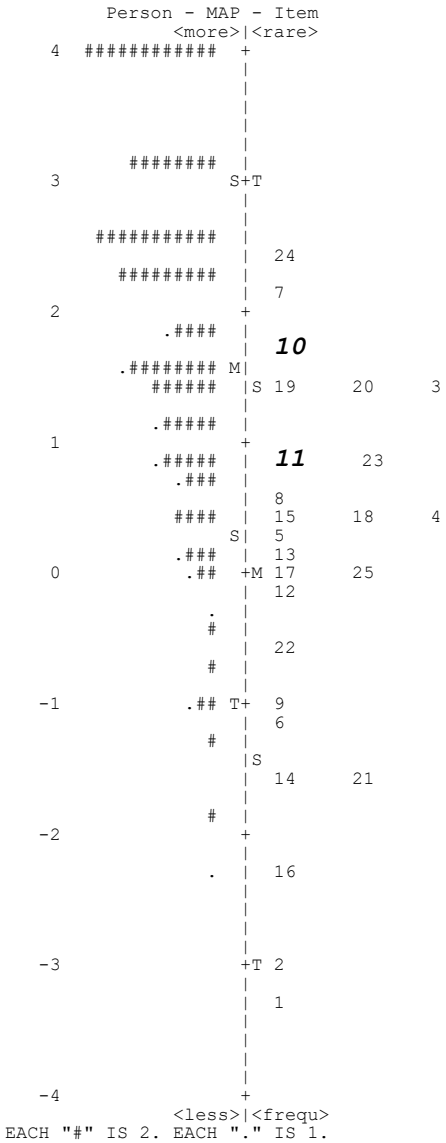

Figure 4. Person/Item Map for the 2016 GF

For the ***Gap-Fill Test***, the following problems were identified: The wording surrounding Items 10 and 11 was somewhat unclear. These two items were somewhat difficult (see bold-italics in Figure 4). The only G-F revision was as follows: Because we were unwilling to scrap the entire test at this time and start afresh, we made only cosmetic changes by slightly rewording the last two lines of the passage to make items 10 and 11 clearer. This clarification had the effect of making these two items easier (compare their positions in Figures 4 here & 7 below).

ANALYSIS AND RESULTS FOR REVISED 2017 ELIPT

2017 Academic Listening Test (ALT)

2017 ALT overall results. Table 4 presents the descriptive statistics for the ALT taken by 169 students during the Spring and Fall Semesters in 2017 (n = 40 and n = 129, respectively). The raw scores for the ALT are approximately normally distributed with the mean, the median, and the mode clustering around about the same value (M = 19.83, SD = 5.05). Table 5 provides a summary of the item-level performance on the test. All items on the ALT are functioning well in terms of their difficulty (neither too easy nor too difficult), but some items still may not be separating the learners' academic listening abilities very well (as shown by the minimum item discrimination value of .02 in Table 5). The internal consistency reliability for the test is .71 as measured by K-R20, a special case of Cronbach's α for dichotomous score, and the standard error of measurement (*SEM*) is 2.74.

Table 4
Descriptive Statistics for ALT for Examinees (n = 169) and Items (k = 36)

Min	Max	Median	Mean	Mode	*SD*	Variance	Skewness	Kurtosis
8	32	20	19.83	18	5.05	25.46	.13	-.49

Table 5
Item Analysis and Reliability Statistics for ALT

Item Facility			Item Discrimination			Reliability		
Min	*Mean*	*Max*	*Min*	*Mean*	*Max*	*K-R20*	*K-R21*	*SEM*
.27	.55	.83	.02	.32	.60	.71	.67	2.74

Table 6 displays Rasch analysis statistics for the ALT measured in logits, which can be used to compare test-takers' performance and items' difficulty on the same interval scale.

Table 6
Rasch Logit Descriptive Statistics of ALT for Persons (Examinees) and (Test) Items*

Facets	Mean	*SD*	Min	Max	RMSE	Separation Index	Reliability
Examinees	.26	.70	-1.41	2.30	.37	1.57	.71
Items	.00	.77	-1.47	1.34	.17	4.32	.95

*Note: Bond & Fox (2006) accompanying *Winsteps* software was used to generate these results.

The score ranges for the examinees and items facets are approximately the same suggesting a good match between the test difficulty and test-takers' ability. However, the examinees' abilities (M = .26, SD = .70) are slightly higher than the item difficulties (M = .00, SD = .77) suggesting that more difficult items may be needed for this test population. The person/item map for the 2017 ALT, which is shown in Figure 5, illustrates this observation.

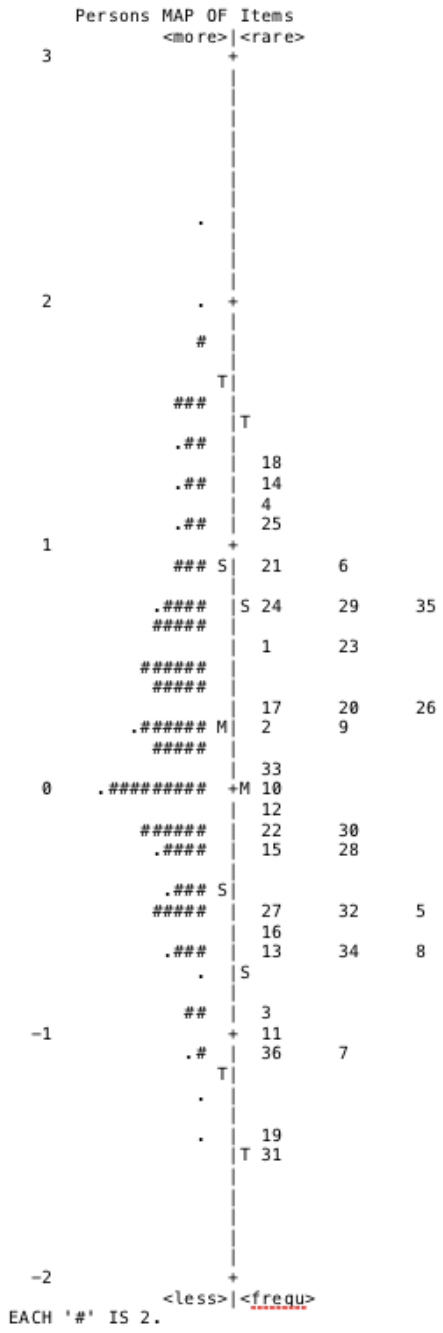

Figure 5. Person/Item Map for the 2017 ALT

2017 ALT item analysis. This section focuses on the item analysis for the ALT. Item facility (IF), item discrimination (ID), and distractor indexes (DI) were analyzed using TAP software (Brooks, 2016). Additional information from the Rasch analysis was also used to examine the flagged items. Overall, most of the items are performing well with desirable IF values (between .30-.70), IDs (above .30), DI (even distractors), and model fit index (between -2 and 2).

Table 7 displays the item analysis for the 2017 ALT. Fourteen items were flagged for closer examination. However, only six items need cross-check using both classical test theory and the Rasch model. Of those six items, only item #3 problematic because the answer in the key (C) is too attractive while another distractor (A) discriminated better than the correct answer. This is also a misfitting item from the Rasch model (z = 1.71).

Table 7
Item Analysis for 2017 ALT

Item	Key	Correct	IF	ID	Point-Biserial Corr	Pt-Biser (Adjusted)	DI	Measure	zInfit Zstd
1	1	71	.42	.50	.42	.33			
2	2	85	.50	.36	.28	.18			
3	3	**125**	.74	**.01**	**.03**	**-.06**	**A***	**-0.9**	**1.71**
4	2	52	.31	.24	.21	.12			
5	3	111	.66	.20	.20	.11			
6	4	61	.36	.48	.44	.36			
7	4	130	.77	.38	.34	.26			
8	2	119	.70	.33	.31	.23			
9	4	83	.49	.60	.49	.41			
10	3	94	.56	.17	.14	.04		.00	2.4
11	1	127	.75	.32	.33	.25			
12	2	97	.57	.26	.21	.11			
13	1	118	.70	.44	.44	.36			
14	4	49	.29	.18	.11	.03		1.24	1.63
15	1	102	.60	.42	.36	.27			
16	2	115	.68	.33	.30	.21			
17	4	81	.48	.18	.14	.04		.34	2.6
18	2	46	.27	.14	.17	.08			
19	2	138	.82	.23	.27	.20			
20	4	81	.48	.07	.16	.07		.34	2.13
21	3	60	.36	.28	.33	.24			
22	3	99	.59	.46	.40	.32			
23	2	71	.42	.39	.30	.20			
24	3	65	.38	.40	.38	.29			
25	3	53	.31	.53	.48	.40			
26	1	82	.49	.45	.42	.34			
27	4	112	.66	.34	.30	.21			
28	1	102	.60	.30	.30	.21			
29	2	67	.40	.50	.42	.34			
30	1	101	.60	.51	.46	.37			
31	1	140	.83	.31	.30	.23			
32	3	111	.66	.42	.35	.26			
33	3	90	.53	.25	.24	.15			
34	4	118	.70	.35	.34	.26			
35	3	65	.38	.09	.10	.01		.77	2.6
36	4	130	.77	.23	.25	.17			

*Indicates that this distractor discriminated better than the correct answer

2017 Reading Comprehension Test (RCT)

2017 RCT overall results. Table 8 provides descriptive statistics for the RCT which was taken by 173 students during the spring and fall semester 2017 (n = 42 and n = 131, respectively). Raw scores of the test appear to be somewhat negatively skewed (-.22) suggesting that the test may be a bit easy for the test-taker population. This observation is corroborated in Table 9 by the mean item difficulty of .58 and maximum item difficulty of .83. Nonetheless, the internal consistency reliability turned out to be .86, indicating the test is doing a good job of consistently separating the examinees. Some items need close inspection as indicated by the *IF* minimum of .02, which indicates an item that is very difficult. Also, the minimum *ID* of -.04 indicates at least one item is not discriminating well between the high scorers and low scorers.

Table 8
Descriptive Statistics of 2017 RCT for Examinees (n = 173) and Items (k = 50)

Min	Max	Median	Mean	Mode	*SD*	Variance	Skewness	Kurtosis
8	46	30	29.42	30	8.33	69.43	-.22	-.63

Table 9
Item Analysis and Reliability Statistics for 2017 RCT

Item Facility			Item Discrimination			Reliability		
Min	*Mean*	*Max*	*Min*	*Mean*	*Max*	*KR20*	*KR21*	*SEM*
.02	.58	.83	-.04	.40	.73	.86	.84	3.09

Table 10 presents summary of Rasch statistics for the RCT. The items (M = .00, SD = .96) appear to be a bit easy for the examinees (M = .42, SD = .89). The separation reliability is .86 and .96, respectively for persons and items.

Table 10
Rasch Logit Descriptive Statistics of RCT for Persons (n = 173) and Items (k = 50)*

Facet	Mean	*SD*	Min	Max	RMSE	Separation Index	Reliability
Examinees	.42	.89	-1.89	2.83	.34	2.44	.86
Test Items	.00	.96	-1.42	4.30	.19	5.06	.96

*Note: Bond & Fox (2006) accompanying *Winsteps* software was used to generate these results.

The person/item map generated by Rasch measurement (shown in Figure 6) provides an overview of the score distribution and the match between item difficulty and person ability. Many items are at the similar difficulty level as displayed by their overlapping (e.g., items 25,

34, 35, 46, 48 or items 13, 14, 18, 4, 41, 50). The spread of item difficulty is needed for the test to perform well. Obviously, item 43 is too difficult and needs further examination.

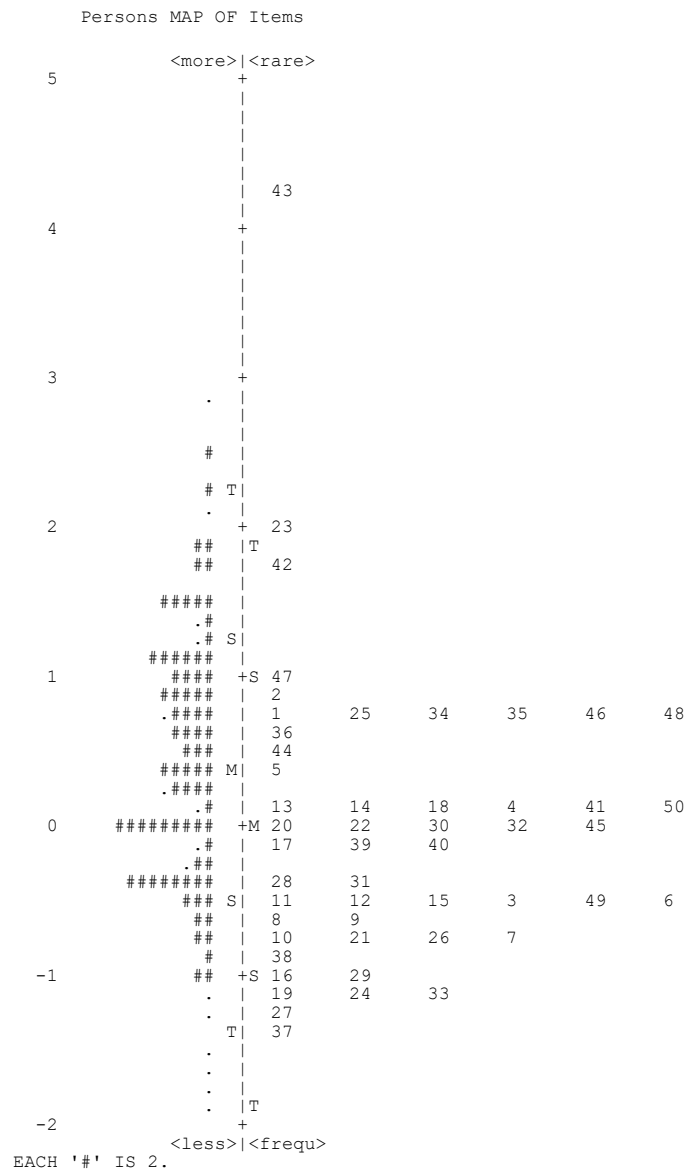

Figure 6. Person/Item Map for the 2017 RCT

2017 RCT item analysis. This section focuses on item analysis for the 2017 RCT. Table 11 shows the item facility (IF), item discrimination (ID), and distractor indexes (DI) that were analyzed for the 2017 RCT using TAP software (Brooks, 2016). Additional information from the Rasch analysis was also used to examine the flagged items. Overall, most of the items in both tests were performing well with desirable IF values between .30 and .70, ID values above .30, DI values indicating even distractors, and model fit indexes within the -2 to 2 range. However, the analyses did indicate that three items were potentially problematic (item #23, #43, #47) as defined by the IF, ID, Point-Biserial correlation, DI, and model fit information (see Table 11).

Table 11
Item Analysis for the RCT

Item	Key	Correct	IF	ID	Point-Biserial Corr	Pt-Biser Adjusted)	DI	Measure	t Infit Zstd
1	4	73	.42	.53	.44	.39			
2	2	70	.40	.74	.55	.51		.87	**-2.91**
3	3	118	.68	.64	.53	.49			
4	1	96	.55	.58	.49	.44			
5	3	86	.50	.69	.52	.48		.43	**-2.32**
6	1	119	.69	.54	.45	.41			
7	1	126	.73	.51	.45	.41			
8	3	121	.70	.33	.35	.30			
9	4	122	.71	.49	.50	.46			
10	1	126	.73	.14	.21	.16			
11	3	119	.69	.25	.25	.20			
12	2	119	.69	.41	.35	.30			
13	2	99	.57	.38	.36	.31			
14	3	97	.56	.48	.40	.34			
15	4	118	.68	.47	.41	.36			
16	3	133	.77	.34	.35	.30			
17	3	104	.60	.46	.41	.36			
18	3	95	.55	.34	.24	.18			
19	4	136	.79	.47	.48	.44			
20	2	101	.58	.31	.28	.22			
21	3	126	.73	.33	.28	.23			
22	1	100	.58	.33	.27	.21			
23	1	34	**.20**	**.09**	**.12**	.07	D*	2.05	1.51
24	4	137	.79	.32	.37	.32			
25	1	76	.44	.32	.29	.23			
26	3	125	.72	.37	.35	.30			
27	3	139	.80	.34	.37	.33			
28	2	113	.65	.45	.45	.40			
29	1	133	.77	.51	.45	.40			
30	2	101	.58	.21	.24	.18			
31	3	115	.66	.39	.30	.25			
32	4	100	.58	.32	.30	.25			
33	1	137	.79	.37	.33	.29			
34	2	75	.43	.28	.29	.24			

35	3	75	.43	.44	.37	.32			
36	2	81	.47	.47	.34	.28			
37	4	144	.83	.26	.28	.23			
38	2	131	.76	.47	.39	.35			
39	4	106	.61	.45	.39	.34			
40	4	108	.62	.46	.34	.29			
41	2	96	.55	.62	.48	.44			
42	4	41	.24	.46	.45	.41			
43	1	5	**.03**	**-.04**	-.10	-.12	B*	4.3	0.34
44	1	84	.49	.67	.47	.42			
45	3	101	.58	.58	.48	.43			
46	2	74	.43	.28	.21	.15		.76	**2.69**
47	2	65	.38	**.02**	**-.02**	-.08	D*	1.01	**5.1**
48	2	73	.42	.45	.39	.33			
49	1	119	.69	.52	.41	.37			
50	4	99	.57	.55	.45	.40			

*Indicates that this distractor discriminated better than the correct answer

Gap-Filling Test (GF)

Tables 12 and 13 present statistics for the raw scores and the logit scores of the gap-filling test. The raw scores from 131 test-takers (Fall 2017[1] only) are negatively skewed (-1.46) suggesting that the test is relatively easy for the current examinee population. The information from the Rasch analysis confirms this observation in that 10 test-takers got all items correct (extreme persons in Rasch). Even with those 10 examinees removed from the analysis, the mean logit for person ability is 1.82 (*SD* = 1.38) in comparison with the item difficulty (*M* = .00, *SD* = 1.22). The reported separation reliability for both examinees and test items are .75 and .94, respectively.

Table 12
Descriptive Statistics of GF for Examinees (n = 131)

Min	Max	Median	Mean	*SD*	Variance	Skewness	Kurtosis	KR-20
2	25	21	19.70	4.54	20.66	-1.46	2.82	.85

Table 13
Rasch Logit Descriptive Statistics of GF for Persons (n = 131) and Items (k = 25)*

Facet	Mean	*SD*	Min	Max	RMSE	Separation Index	Reliability
Examinees	2.08	1.61	-2.96	5.21	.84	1.64	.73
Examinees**	1.82	1.38	-2.96	3.90	.69	1.75	.75
Test Items	.00	1.22	-2.00	2.90	.30	3.95	.94

*Note: Bond & Fox (2006) accompanying *Winsteps* software was used to generate these results.
** Non-extreme persons (*n* = 121)

[1] Data for the Spring Administration was not available.

Notice in Figure 7 that the person/item map indicates a continuing general mismatch between person abilities and test difficulties. Most items are easy for the most test-takers. Also notice that items 10 and 11 are considerably easier in this administration of the test than they were in 2016. Rewording them and making the clearer apparently worked because it made them easier here (see bold italics in Figure 7) than they were in the previous version (see Figure 4 above). Also note that only items 7 and 24 seem to function well at separating the high-level examnees' abilities. The GF definitely needs to be reworked—perhaps replaced entirely.

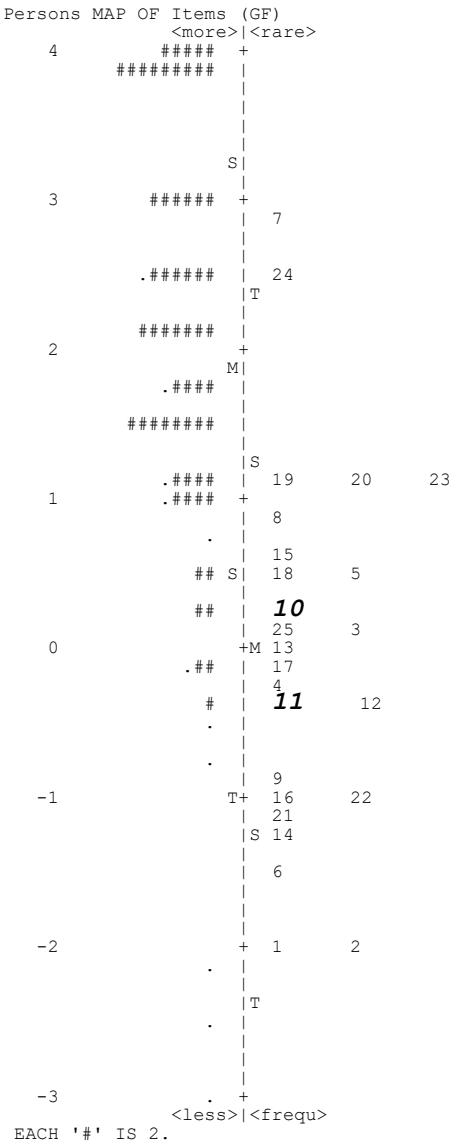

Figure 7. Person/Item Map for the 2017 GF Test

The Writing Test

Table 14 shows the descriptive statistics for the 110 writing test-takers during the Fall.

Table 14
Descriptive Statistics of WTG for Examinees (n = 110)

Min	Max	Median	Mean	*SD*	Variance	KR-20
20	33	26.33	26.55	3.03	9.18	.91

As shown in Table 15, each essay was read by three different raters and given scores using the four-category rubric. Table 15 presents the descriptive statistics for the six raters and 122 examinees in this study. Six statistics are listed in the first column (*N* = the *number* of essays read by each rater; the *mean* in this case is equivalent to the arithmetic average; the *median* is that point that divides the scores 50/50; the *SD* = the *standard deviation* is an indicator of the dispersion of the scores because it is a sort of average of the distances of scores from the mean; the *min* is the minimum or lowest score given; and the *max* is the maximum or highest score. Given that the three readings are slightly different in mean scores and the standard diviations, closer examination of variation between raters may be needed.

Table 15
Descriptive Statistics for Three-Reading Totals and Categories

Statistic	**Mean**	**Median**	***SD***	**Min**	**Max**
Rg1 Con	7.12	7.00	1.20	5.00	9.00
Rg1 Org	6.87	7.00	1.13	5.00	9.00
Rg1 Voc	7.17	7.00	1.20	5.00	10.00
Rg1 Gram	7.33	7.00	1.14	5.00	9.00
Rd1 Total	**28.49**	**28.00**	**4.26**	**20.00**	**37.00**
Rg2 Con	7.35	7.00	1.13	5.00	10.00
Rg2 Org	7.19	7.00	1.11	5.00	9.00
Rg2 Voc	7.25	7.00	1.11	5.00	10.00
Rg2 Gram	7.21	7.00	1.06	5.00	10.00
Rd2 Total	**28.99**	**29.00**	**3.87**	**20.00**	**39.00**
Rg3 Con	7.7	7.50	1.19	5.00	10.00
Rg3 Org	7.7	7.00	1.22	5.00	10.00
Rg3 Voc	8.4	7.50	1.27	5.00	10.00
Rg3 Gram	8	7.00	1.26	5.00	10.00
Rd3 Total	**31.80**	**30.00**	**4.44**	**20.00**	**39.00**

Table 16
Rasch Logit Descriptive Statistics of the Writing Test*

Facet	Mean	*SD*	Min	Max	RMSE	Separation Index	Reliability
Examinees*	-.40	1.21	-3.65	2.23	.35	3.28	.91
Readings	.00	.19	-.25	.22	.60	3.90	.94
Categories	.00	.23	-.15	.35	.07	3.35	.92

* Note: Bond & Fox (2006) accompanying *Facets* software was used to generate these results.
** Non-extreme persons (n = 121)

The overall Rasch Facets analysis results for the WTG are shown in Table 16. And, the vertical yardstick for these analyses is shown in Figure 8 for three different facets (Examinees, Readings, & Categories) on the same scale. Again, the three Readings differ in their severity as described above. Of the four categories in the rubric, essay organization is most severe in terms of its difficulty in comparison with the three other domains (vocabulary, content, & grammar). Overall, the writing test is performing well as indicated by the good match between different facets on the same scale (examinee, ratings, categories).

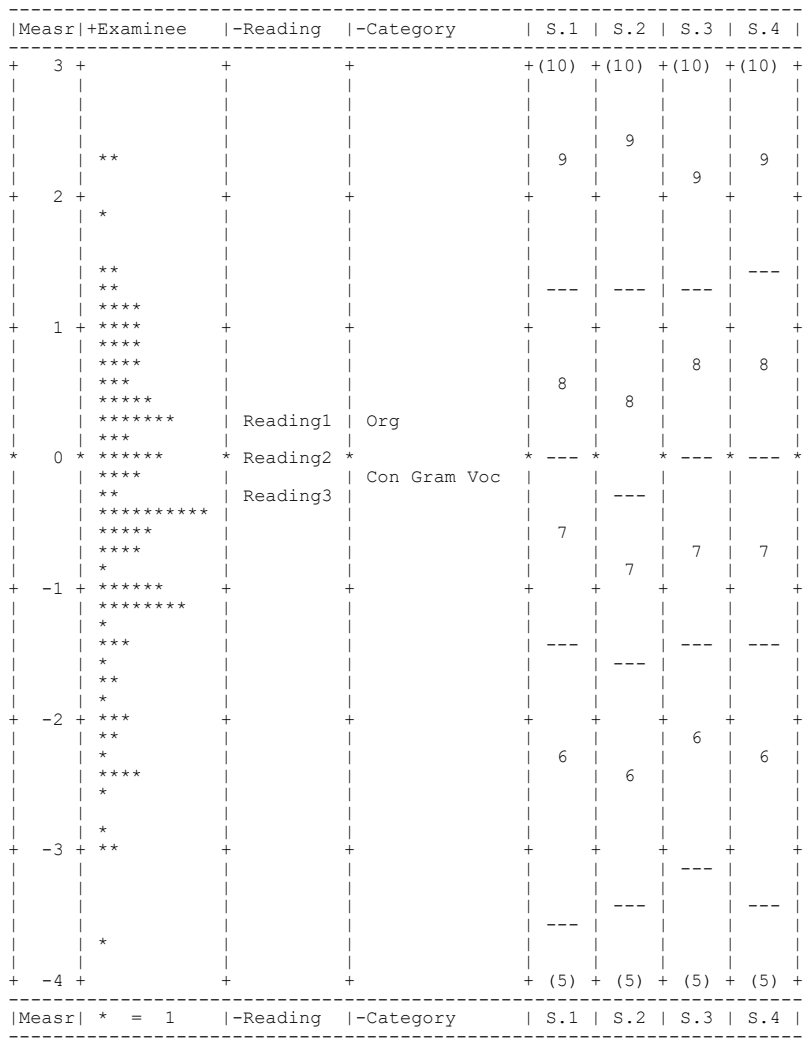

Figure 8. Rasch FACETs Analysis Vertical Ruler for the 2017 WTG Test

CONCLUSIONS

This report has explained how the 2016 ELIPT was revised and then provided an overview of the test score distribution, reliability, and item analyses for the 2017 ELIPT. The revisions and analyses were carried out for the ALT and RCT using both the classical test theory and Rasch model analyses. The Writing (WTG) and the Gap-Fill (GF) tests were also examined using the Rasch analysis.

Table 17 summarizes and compares the descriptive statistics for the original 2016 and revised 2017 versions of the ELIPT. Notice that the reliability estimates in the bottom row of the table are generally higher for the 2017 ELIPT (except for the WTG, which stayed the same at .91).

This increase in reliability may indicate that this test-revision project improved the ELI testing and decision-making process generally. While it is possible that the ability levels of the students in 2017 were more widely disbursed than those in 2016, which in turn would have led to higher reliability, a quick examination of the numbers of items, means, medians, standard deviations, minimum and maximum values and ranges for the two years shown in Table 17 does not systematically support that interpretation. However, the fact that the DCT reliability also improved from 2016 to 2017, even though only cosmetic changes were made to that test, is worth noting.

Table 17
Comparing Descriptive Statistics for ELIPT 2016

Statistic	**2016 (Fall & Spring)**					**2017 (Fall & Spring)**				
	ALT	**DCT***	RCT	GF	WTG*** (Exper.)	ALT	**DCT***	RCT	GF (F17)	WTG (F17)
N	182	**182**	182	182	122	169	**169**	173	131	110
K	35	**50**	50	25	40	36	**50**	50	25	40
Mean	19.01	**30.40**	28.33	18.59	27.39	19.83	**31.59**	29.42	19.70	26.55
Median	19	**31**	29	19	27	20	**32**	30	21	26.33
SD	4.59	**8.34**	7.98	4.76	3.08	5.05	**9.50**	8.33	4.54	3.03
Max	31	**49**	45	25	21	32	**50**	46	25	33
Min	6	**10**	9	4	37	8	**5**	8	2	20
Range	26	**40**	37	22	17	25	**46**	39	24	14
Reliability	.67	**.73****	.68	.77	.91	.71	**.89****	.86	.85	.91

*Bold DCT only cosmetic changes made in test
**K-R21 = very rough estimate
***See Brown, Hsu, Harsch (2017)

Based on the analysis in this project, we recommend the following general changes be made to further improve the ELIPT moving forward:

1. Qualitatively examine and revise the items on the ALT and RCT.
2. Revise and develop an entirely new form for the Gap-Fill test
3. Closely examine the rater variability for the writing exam and develop the rater training program if applicable.

REFERENCES

Brown, J. D. Hsu, W.-L., & Harsch, K. (2017). 2017 ELIPT writing test revision project. *Second Language Studies, 35*(2), 1-29.

Phung, H. V. (2017). *ELIPT 2017 test & item analysis report.* English Language Institute, Department of Second Language Studies, University of Hawai‘i at Mānoa, Honolulu, HI.

Bond, T. G., & Fox, C. M. (2006). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.

Brooks, G. P. (last update 2016). *TAP: Test analysis program.* Available online at https://people.ohio.edu/brooksg/#TAP.