The Revision and Development of the HELP Placement Test

Wenpei Long
Ann Johnstun

Department of Second Language Studies
University of Hawaiʻi at Mānoa

As advised by J.D. Brown
SLS 699, Directed Reading

June, 2009

Introduction

This study focuses on the placement test at the Hawaii English Language Program (HELP), an Intensive English Program (IEP). The Director of HELP, Joel Weaver, has identified the test as a weakness in the placement procedures and has invited SLS students to assist in the revision and development of the test. The current placement test is time-consuming and somewhat redundant. There have also been some problems in the placement results in the recent past, mainly that the procedures in proctoring the test are not well-known or readily available. This allows for much variation in the test results as many of the test administrators and assessors are new and untrained. This paper will describe each section of the test, analyze effectiveness of each test for placement purposes, and give recommendations for the future placement test based on our findings. The investigation aims to evaluate the HELP placement test in two ways: through statistical analysis using Classical Testing Theory (CTT) and Rasch analysis to find whether the test itself is functioning to place students correctly, and by evaluating the overall test design, procedures, and logistics to find whether the test is being proctored and scored correctly.

Background

The existing test consists of four parts; the HELP reading test, the HELP writing test, a speaking conversation test, and the Michigan EPT test. All of the tests except for the Michigan EPT have been designed in-house specifically for HELP placement purposes. These tests are scored individually and are considered together in order to place students into one of four levels.

Reading test

The HELP Reading test consists of 40 multiple-choice reading items. Students are given 55 minutes to take the test. The exam was designed to evaluate a student’s reading ability relative to the range of reading materials across the HELP curriculum. Excerpts from actual readings used in HELP courses provided the basis for item construction. Each item consists of a short passage, approximately 20-95 words, which includes a cloze sentence at the end. The final cloze sentence functions as a reading comprehension prompt, and the students must use the information in the passage to best complete the sentence. The item response options are all single-word answers. This test was last revised in 2005 (Johnson, 2006).

Writing test

The writing test is an open response exam, with four optional questions for examinees to answer. Students are asked to choose one topic from the four options and write an essay. They are given 45 minutes to write, and are supplied with as much paper as they need to complete their essay. The writing test responses are then rated by at least two raters in accordance with a rubric developed at HELP to align with course difficulty.

Speaking Conversation

The speaking conversation test consists of three parts: an oral interview, a dictation, and a video listening. The video listening part of the test has recently been replaced by the Michigan EPT listening section because the quality of the video and sound diminished over time which caused the test to be less-reliable. The oral interview is a group interview, with one interlocutor and three or four examinees being tested at one time. There are no existing guidelines for how the group interview should be conducted, therefore new guidelines or a new test may be developed in its place.

Michigan EPT

The Director has suggested that HELP switch to use the Michigan EPT because he is more familiar and feels more comfortable with this test. Subsequently, all continuing students were given the Michigan EPT in the hope that the newly introduced test could be compared and correlated with the old HELP placement test in order to choose the most effective and efficient model.

First Steps in Testing Reform

The initial phase of this project consisted of meeting with HELP's Director to clearly identify our role and duties in the project. In this meeting, all initially agreed that we (the researchers) would complete the following tasks:

1. Analyze the Michigan EPT test results from HELP students, to inform whether this can be used as an effective placement exam at HELP
2. Evaluate and reform the reading and writing tests
3. Develop a new speaking test

During the imputing, processing and analyzing testing data, we found many errors associated with a lack of written test administration and scoring procedures. Therefore, developing new test administrating and scoring procedures were also included in the agenda.

4. Revise written test administration and scoring procedures

Participants

We obtained data from 47 new and continuing students who took the Michigan EPT, reading, and writing tests. These students vary in age, language background, and time spent in English-speaking environments. Of the 47, students were from Japan, Korea, China, Thailand, Peru, Colombia, Brazil, and Turkey. They ranged in age from 18-50+, and some had been in the US only one day while others have been here for over ten years.

Method

Contrary to the classical Item Response Theory (IRT), we will employ the Rasch model (Bond & Fox, 2007) for measurement. The Rasch model incorporates an algorithm that expresses the probabilistic expectations of item and person performances (Wright & Stone, 1979, cited in Bond & Fox, 2007, p. 37). In the present study, we will only consider the one-parameter Rasch model which includes two important variables of the item difficulty and the human ability. Each item difficulty and person ability is integrated and estimated on a logit (log odds unit) scale, an interval scale (Bond & Fox, 2007, p. 41). Two key propositions are drawn to illustrate the basic principles of the Rasch model:

1. Persons who are more able or more developed have a greater likelihood of correctly answering all the items in the observation schedule
2. Easier items are more likely to be answered or reached correctly by all persons.

(Bond & Fox, 2007, p. 37)

Findings

Michigan EPT test and HELP Reading Test

In this section, our analysis focuses on the Michigan EPT test and the HELP Reading test. As stated previously, HELP started to employ Michigan EPT test as part of placement at term Fall II, 2008. Due to a lack of data and analysis on how well the EPT performed to discriminate students, the HELP reading test was kept as to provide adequate information for placement decisions. However, each test takes one hour to finish, with the whole placement test taking almost four hours to administer. This is both time and labor consuming, and the HELP administration felt that we could shorten the test while maintaining reliable results. A systematic analysis was conducted to solve this problem. Several research questions were brought up as guidance for analysis.

1. How well do the listening and reading sub-sections of the EPT perform to discriminate HELP students?

2. How well does the EPT spread out HELP students overall?
3. What is the correlation between the EPT and the HELP Reading test? Can the EPT Reading section substitute the HELP Reading test to place the students in terms of their reading abilities?

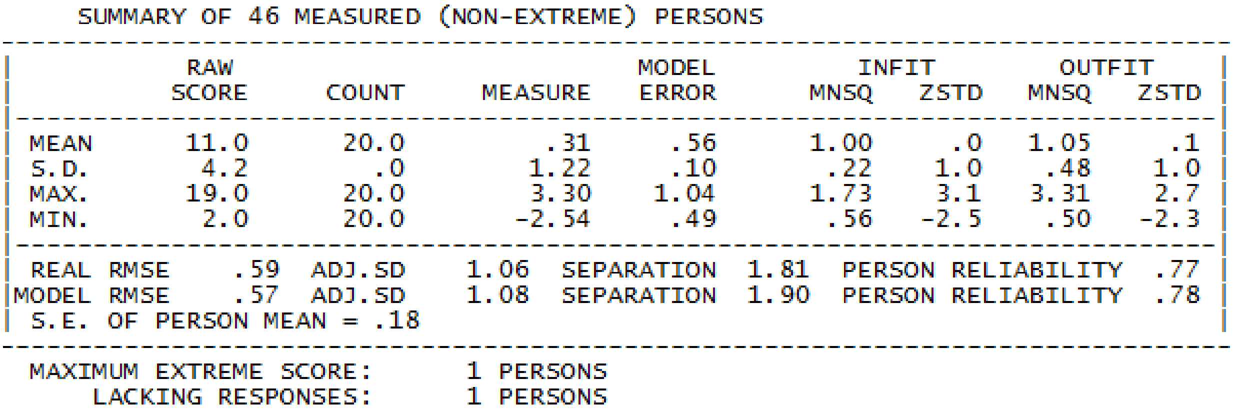
The Michigan EPT test consists of 100 multiple-choice items, with 20 listening items, 30 vocabulary items, 30 grammar items and 20 reading items. 47 students at HELP took the Michigan EPT test. We transferred the testing results into dichotomous data with value 1 and 0 in order to run WINSTPES (Linacre, 2008), the Rasch model computer program.

1. How well did the Listening and Reading sections perform to discriminate HELP students?

Table 1 provides the summary of person measurement in the Listening section. *Extreme*, in contrast of *non-extreme,* summarizes persons with extreme scores (includes zero and perfect scores) (WINSTEPS Manual, 2008). Non-extreme, therefore, summarizes persons without extreme scores. *Measure* is the estimated logit of person ability in relation to the item. *Infit* is an information-weighted sum, giving relatively more weight to the performances of persons closer to the item value. Infit MNSQ, or Infit mean square, is the model variance, with an expected value of 1, and Infit Zstd, or Infit Z, is the model standard deviation, with and expected value of 0. The person reliability is equivalent to the traditional “test” reliability. According to the tentative guideline from the Winsteps Manual, the test with the person reliability of .90 can discriminate proximately 3 or 4 levels, person reliability of .80, 2 or 3 levels, person reliability of .50, 1 or 2 levels. Based on the simple explanation of the analysis reports produced from the Winsteps, we can see that the reliability of this test is quite high, with a value of .77 at the lower bound and .78 at the upper bound. Cronbach Alpha (KR-20) is calculated .81. We can find out that the person ability logit value ranges from the minimum of -2.54 to the maximum of 3.30, with a mean person ability logit value of .31, indicating that the difficulty level of the Listening section is quite fit the ability of the majority of students taking the test. The Infit mean square is estimated 1.00 and the Infit Z is estimated .0 with a standard deviation of 1, which can be interpreted as perfect fit to the Rasch model. In other words, the students performed quite normal in the listening section. Only one person got the maximum extreme score.

Table 1

The Summary of Person Measurement in the Listening Section (Non-Extreme)

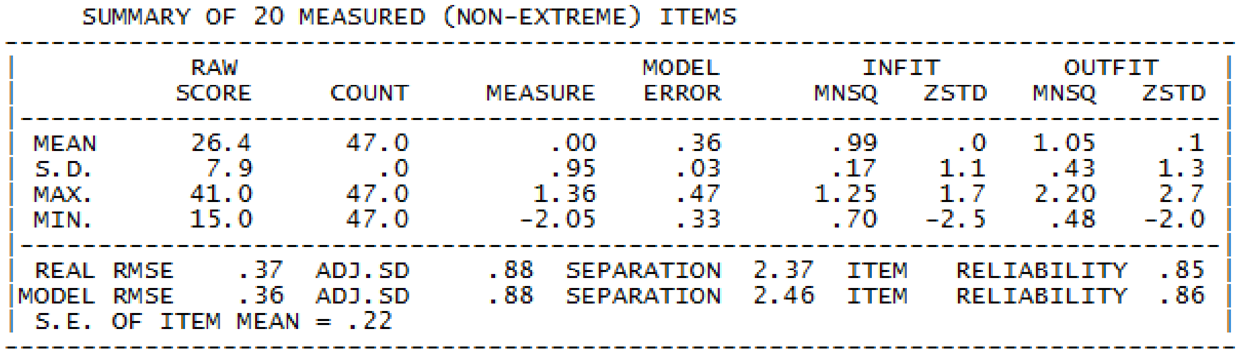


```
    SUMMARY OF 46 MEASURED (NON-EXTREME) PERSONS
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT      |
|         SCORE     COUNT    MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN     11.0      20.0        .31      .56      1.00     .0   1.05     .1  |
| S.D.      4.2        .0       1.22      .10       .22    1.0    .48    1.0  |
| MAX.     19.0      20.0       3.30     1.04      1.73    3.1   3.31    2.7  |
| MIN.      2.0      20.0      -2.54      .49       .56   -2.5    .50   -2.3  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .59  ADJ.SD    1.06  SEPARATION  1.81  PERSON RELIABILITY  .77 |
|MODEL RMSE    .57  ADJ.SD    1.08  SEPARATION  1.90  PERSON RELIABILITY  .78 |
| S.E. OF PERSON MEAN = .18                                                   |
-------------------------------------------------------------------------------
 MAXIMUM EXTREME SCORE:      1 PERSONS
     LACKING RESPONSES:      1 PERSONS
```

Table 2 reports the summary of item measurement in the listening section. Item reliability, different from the person reliability, has no traditional equivalent. Item reliability and item separation refer to the ability of the test to define a distinction hierarchy of item along the measured variable (Bond & Fox, 2007, p. 60). The higher the item reliability and item separation are calculated, the more confident we can claim that the test could be applied to other suitable samples. Therefore, the item reliability of the Listening section is estimated .85, high enough for us to be confident that the order of item estimates in this test could be replicated to other samples for whom it is suitable. The mean measure of item difficulty is estimated .00, with a maximum logit value of 1.36 and a minimum logit value of -2.05. However, compared with the range of the person ability measure (-2.54 - 3.30), we can notice that the range of the item difficulty measure (-2.05 - 1.36) in the listening section is not large enough to measure and discriminate the students with higher ability. This can be shown in the item map in a more visual way (see Figure 1). Figure 1 also shows that the clusters of items with similar difficulty measures do not match the clusters of persons with similar ability measures, which reveals that the items in the Listening section may not function well to discriminate according to its cut-point.

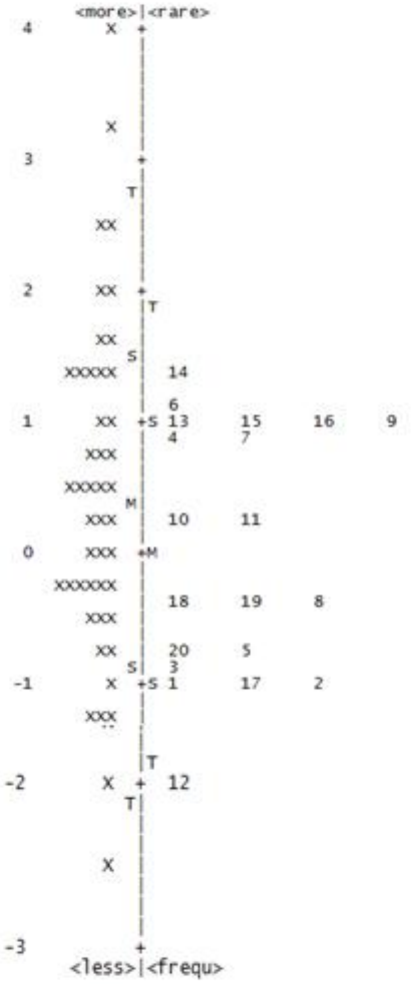
Table 2

The Summary of Item Measurement in the Listening Section



```
     SUMMARY OF 20 MEASURED (NON-EXTREME) ITEMS
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT     |
|          SCORE      COUNT     MEASURE    ERROR      MNSQ    ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       26.4      47.0        .00      .36        .99      .0   1.05     .1 |
| S.D.        7.9        .0        .95      .03        .17     1.1    .43    1.3 |
| MAX.       41.0      47.0       1.36      .47       1.25     1.7   2.20    2.7 |
| MIN.       15.0      47.0      -2.05      .33        .70    -2.5    .48   -2.0 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .37  ADJ.SD     .88  SEPARATION  2.37  ITEM   RELIABILITY  .85 |
|MODEL RMSE    .36  ADJ.SD     .88  SEPARATION  2.46  ITEM   RELIABILITY  .86 |
| S.E. OF ITEM MEAN = .22                                                     |
-------------------------------------------------------------------------------
```

Person -map- Item



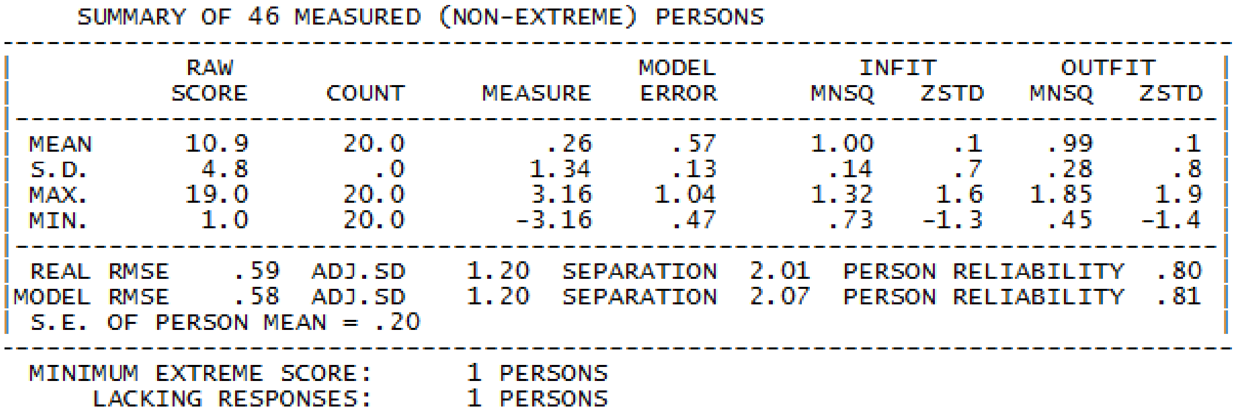
```
     <more>|<rare>
 4       X  +
            |
            |
            |
         X  |
 3          +
           T|
        XX  |
            |
 2      XX  +
            |T
        XX  |
           S|
     XXXXX  |  14
            |  6
 1      XX  +S 13     15     16     9
            |  4      7
       XXX  |
     XXXXX  |
           M|
       XXX  |  10     11
 0     XXX  +M
    XXXXXX  |
            |  18     19     8
       XXX  |
        XX  |  20     5
           S|  3
-1       X  +S 1      17     2
       XXX  |
            |
            |T
-2       X  +  12
           T|
            |
         X  |
            |
            |
-3          +
     <less>|<frequ>
```

Figure 1: Item Map of Listening Section

Table 3 provides a detailed summary of person measurement in the reading section of Michigan EPT test. The person reliability is estimated .80 at the lower bound and .81 at the upper bound, indicating high test reliability. The mean person ability measure is .26, which means that the difficulty level of the HELP reading section quite fit the students' reading abilities. The maximum person ability is measured 3.16 and the minimum is -3.16, with a standard deviation of 1.34, demonstrating a widespread person ability range. The Infit mean square is calculated 1.00 and the Infit Z is calculated .10 with a standard deviation of .70, indicating that the HELP students performed quite normal too in the reading section, except one person with a minimum extreme score.

Table 3

The Summary of Person Measurement in the Reading Section



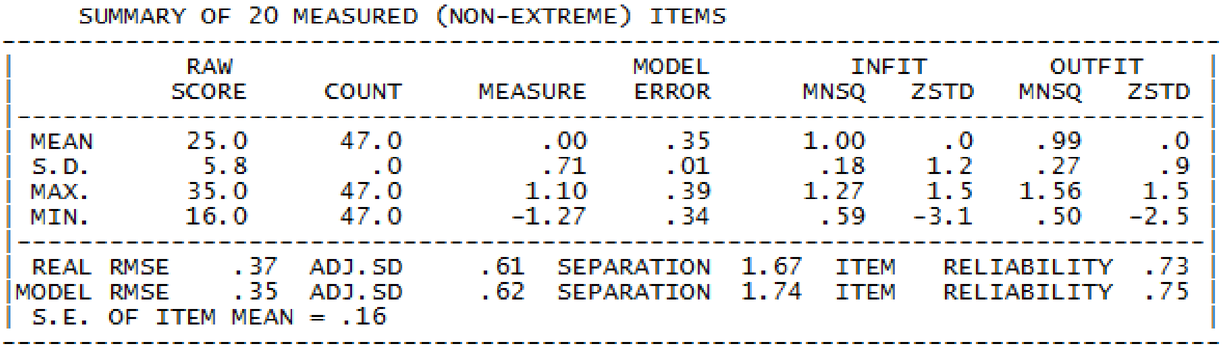
```
     SUMMARY OF 46 MEASURED (NON-EXTREME) PERSONS
-------------------------------------------------------------------------------
|           RAW                          MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      10.9      20.0         .26     .57      1.00     .1    .99     .1 |
| S.D.       4.8        .0        1.34     .13       .14     .7    .28     .8 |
| MAX.      19.0      20.0        3.16    1.04      1.32    1.6   1.85    1.9 |
| MIN.       1.0      20.0       -3.16     .47       .73   -1.3    .45   -1.4 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .59  ADJ.SD    1.20  SEPARATION  2.01  PERSON RELIABILITY  .80 |
|MODEL RMSE    .58  ADJ.SD    1.20  SEPARATION  2.07  PERSON RELIABILITY  .81 |
| S.E. OF PERSON MEAN = .20                                                   |
-------------------------------------------------------------------------------
  MINIMUM EXTREME SCORE:      1 PERSONS
      LACKING RESPONSES:      1 PERSONS
```

In Table 4, the summary of item measures in the Reading section of EPT was reported. The mean measure of item difficulty is estimated .00. The Infit mean square is calculated 1.00, with the Infit Z of .00, which indicates that the items in the Reading section perfectly fit the Rasch model. The item reliability is .73 at the lower bound and .75 at the upper bound, meaning this section can also be replicated to other samples. The measure of the item difficulty ranges from a logit value of -1.27 as the minimum and a logit value of 1.10 as the maximum, with a standard deviation of .71. Similar to the Listening section, the magnitude of the item difficulty (-1.27 – 1.10) is not broad enough to measure and discriminate the person ability (-3.16 – 3.16). This also can be shown in the item-person map (see Figure 2).

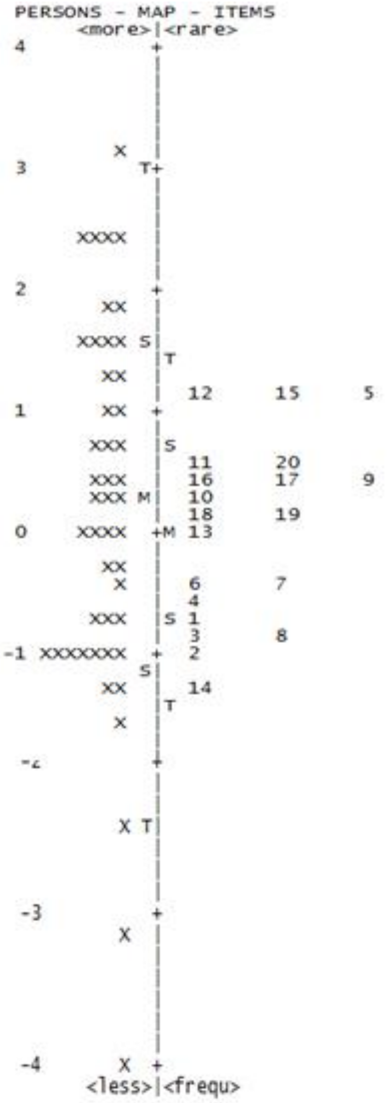
Figure 2 shows that the range of the person ability is much larger than that of the items could measure. This could be resulted from the small range of the item difficulty measures and inadequate number of items in comparison to the sample size. Moreover, we can see that the students with similar person ability measures cluster at a lower logit scale of around -1.0. However, the items with similar difficulty measures cluster at a higher logit scale between 0.0 - +1.0. This means that the item cluster does not match the student cluster. The items may not function well to discriminate students into the levels we expect.

Table 4

The Summary of Item Measurement in the Reading Section



```
     SUMMARY OF 20 MEASURED (NON-EXTREME) ITEMS
-------------------------------------------------------------------------------
|           RAW                               MODEL         INFIT        OUTFIT    |
|           SCORE      COUNT       MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN        25.0       47.0          .00      .35      1.00     .0    .99     .0 |
| S.D.         5.8         .0          .71      .01       .18    1.2    .27     .9 |
| MAX.        35.0       47.0         1.10      .39      1.27    1.5   1.56    1.5 |
| MIN.        16.0       47.0        -1.27      .34       .59   -3.1    .50   -2.5 |
|------------------------------------------------------------------------------|
| REAL RMSE     .37  ADJ.SD      .61  SEPARATION  1.67  ITEM    RELIABILITY  .73 |
|MODEL RMSE     .35  ADJ.SD      .62  SEPARATION  1.74  ITEM    RELIABILITY  .75 |
| S.E. OF ITEM MEAN = .16                                                      |
-------------------------------------------------------------------------------
```



```
 PERSONS - MAP - ITEMS
       <more>|<rare>
 4           +
             |
             |
          X  |
 3          T+
             |
       XXXX  |
             |
 2           +
         XX  |
       XXXX S|
             |T
         XX  |
             |  12      15     5
 1       XX  +
        XXX  |S
             |  11      20
        XXX  |  16      17     9
        XXX M|  10
             |  18      19
 0     XXXX  +M 13
         XX  |
          X  |  6       7
             |  4
        XXX  |S 1
             |  3       8
-1  XXXXXXX  +  2
            S|
         XX  |  14
             |T
          X  |
-2           +
             |
          X T|
             |
-3           +
          X  |
             |
             |
-4        X  +
       <less>|<frequ>
```

Figure 2: Item Map of Reading Section

To conclude, both the Listening and the Reading section have very high reliabilities. Both fit statistics of the person and the item match the Rasch model very well, meaning that the items and persons performed as they are expected. Overall, the Listening and Reading sections perform well to discriminate students and spread them out into levels. However, the range of the item difficulty from both sections is not large enough to measure and discriminate students with extremely high or low person ability. Therefore, in order to function better as to make the placement decision, more items with wider range of difficulty measures are needed in these two sections. For the Listening section, more difficult items are needed. The reading section needs both easier and more difficult items to measure and discriminate the students with a wide range of abilities.

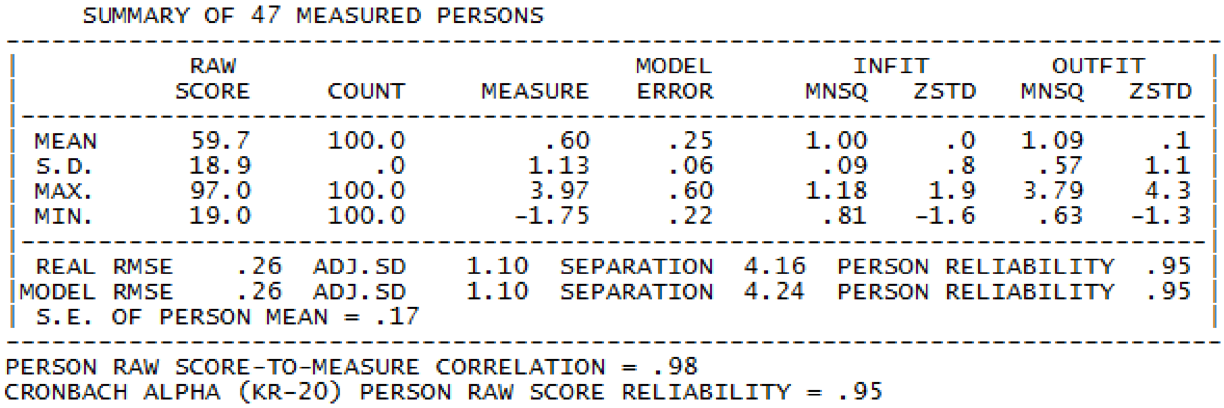
2. How well did the EPT spread out HELP students overall?

In Table 5, we can see that Michigan EPT as an overall test has a really high reliability of .95 at both lower and upper bound. Cronbach Alpha is also estimated .95. The mean measure of person ability is .60. The Infit mean square is 1.00, and the Infit Z is estimated .0 with a standard deviation of .80, which fits well enough to the Rasch model. The measure of the person ability ranges from the logit value of -1.75 to the logit value of +3.97.

Table 6 provides the item measurement of Michigan EPT. The mean measure has a logit value of .00. The Infit mean square is calculated as 1.00. The Infit Z is calculated as .00, with a standard deviation of 1.0, demonstrating perfect fit statistics for the Rasch model, meaning that EPT test functions well to spread out students in to order. The measure of the item difficulty ranges from the logit value of -3.75 to the logit value of +2.73. In relation to the measures of the person ability, some items with low difficulty measure are too easy for all students at HELP. In other words, everybody got these items right and these items could not provide information to discriminate students. However, it is almost in need of easy items that everybody can get right in a certain test.

Table 5

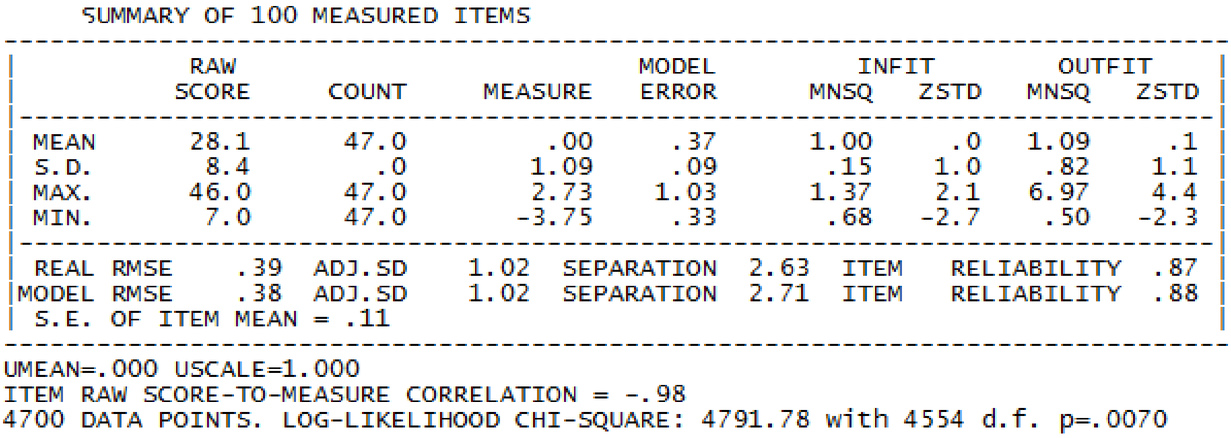
The Summary of Person Measurement of the Michigan EPT



```
     SUMMARY OF 47 MEASURED PERSONS
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT         OUTFIT    |
|          SCORE     COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      59.7     100.0         .60    .25      1.00     .0   1.09     .1  |
| S.D.      18.9        .0        1.13    .06       .09     .8    .57    1.1  |
| MAX.      97.0     100.0        3.97    .60      1.18    1.9   3.79    4.3  |
| MIN.      19.0     100.0       -1.75    .22       .81   -1.6    .63   -1.3  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .26  ADJ.SD    1.10  SEPARATION  4.16  PERSON RELIABILITY  .95 |
|MODEL RMSE    .26  ADJ.SD    1.10  SEPARATION  4.24  PERSON RELIABILITY  .95 |
| S.E. OF PERSON MEAN = .17                                                   |
-------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .95
```

Table 6

The Summary of Item Measurement of Michigan EPT



```
     SUMMARY OF 100 MEASURED ITEMS
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT         OUTFIT    |
|          SCORE     COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      28.1      47.0         .00    .37      1.00     .0   1.09     .1  |
| S.D.       8.4        .0        1.09    .09       .15    1.0    .82    1.1  |
| MAX.      46.0      47.0        2.73   1.03      1.37    2.1   6.97    4.4  |
| MIN.       7.0      47.0       -3.75    .33       .68   -2.7    .50   -2.3  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .39  ADJ.SD    1.02  SEPARATION  2.63  ITEM   RELIABILITY  .87 |
|MODEL RMSE    .38  ADJ.SD    1.02  SEPARATION  2.71  ITEM   RELIABILITY  .88 |
| S.E. OF ITEM MEAN = .11                                                     |
-------------------------------------------------------------------------------
UMEAN=.000 USCALE=1.000
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.98
4700 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 4791.78 with 4554 d.f. p=.0070
```

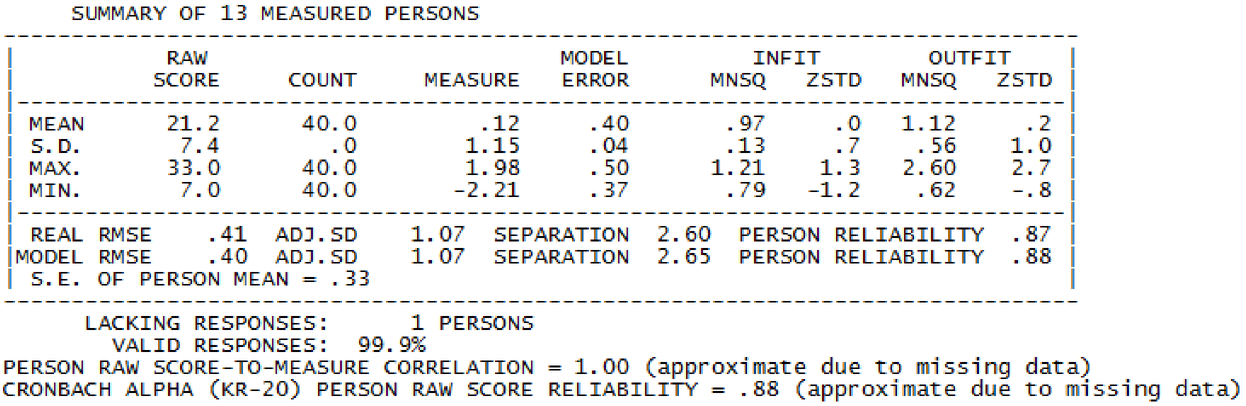
3. What is the correlation between the EPT Reading section and the HELP Reading test? Can the EPT reading section substitute the reading test as to place the students in term of their reading abilities?

The HELP Reading test consists of 40 multiple-choice reading items. The test takes about 55 minutes to finish. We transferred the test results from 13 new students in Term Fall II into dichotomous data of 0 and 1, and ran the transferred data on WINSTPES.

Table 7 shows the detailed report of the person ability measurement. We can see that the HELP Reading has a rather high reliability of .87 at the lower bound and .88 at the higher bound. Cronbach Alpha (KR-20) is also estimated .88, meaning this test is rather reliable. The mean measure of the person ability is .12, indicating that the difficulty level of the test matches the majority of students taking the test. The Infit mean square is calculated .97, very close to the expected value of 1 in the Rasch model. The Infit Z is .0 with a standard deviation of .7, also rather close to the expected value of 1. This means that the difficulty of the test quite fits the ability of the sample size on a whole. Table 8 shows that the item reliability index of the HELP Reading test is .66 at the lower bound and .70 at the higher bound, indicating that the order of item estimates might not be replicated to other samples for whom it is suitable.

Table

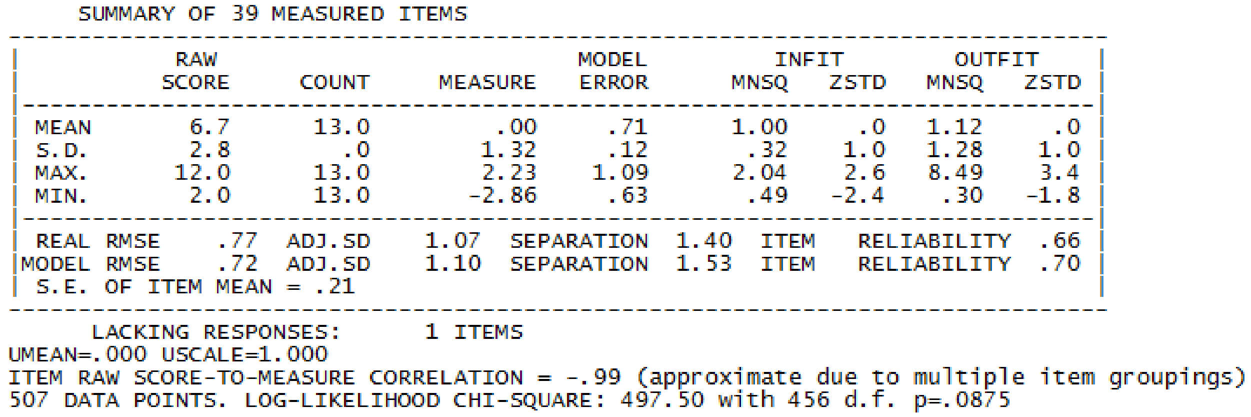
The Summary of Person Measurement of the HELP Reading test



```
     SUMMARY OF 13 MEASURED PERSONS
-------------------------------------------------------------------------------
|           RAW                          MODEL         INFIT         OUTFIT    |
|          SCORE      COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      21.2       40.0         .12     .40       .97     .0   1.12     .2 |
| S.D.       7.4         .0        1.15     .04       .13     .7    .56    1.0 |
| MAX.      33.0       40.0        1.98     .50      1.21    1.3   2.60    2.7 |
| MIN.       7.0       40.0       -2.21     .37       .79   -1.2    .62    -.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .41  ADJ.SD    1.07  SEPARATION  2.60  PERSON RELIABILITY  .87 |
|MODEL RMSE    .40  ADJ.SD    1.07  SEPARATION  2.65  PERSON RELIABILITY  .88 |
| S.E. OF PERSON MEAN = .33                                                   |
-------------------------------------------------------------------------------
      LACKING RESPONSES:      1 PERSONS
        VALID RESPONSES:   99.9%
PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .88 (approximate due to missing data)
```

Table 8

The Summary of Item Measurement of the HELP Reading test



```
     SUMMARY OF 39 MEASURED ITEMS
-------------------------------------------------------------------------------
|           RAW                          MODEL         INFIT         OUTFIT    |
|          SCORE      COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       6.7       13.0         .00     .71      1.00     .0   1.12     .0 |
| S.D.       2.8         .0        1.32     .12       .32    1.0   1.28    1.0 |
| MAX.      12.0       13.0        2.23    1.09      2.04    2.6   8.49    3.4 |
| MIN.       2.0       13.0       -2.86     .63       .49   -2.4    .30   -1.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .77  ADJ.SD    1.07  SEPARATION  1.40  ITEM   RELIABILITY  .66 |
|MODEL RMSE    .72  ADJ.SD    1.10  SEPARATION  1.53  ITEM   RELIABILITY  .70 |
| S.E. OF ITEM MEAN = .21                                                     |
-------------------------------------------------------------------------------
      LACKING RESPONSES:      1 ITEMS
UMEAN=.000 USCALE=1.000
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.99 (approximate due to multiple item groupings)
507 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 497.50 with 456 d.f. p=.0875
```

In order to investigate whether the EPT Reading can substitute the HELP Reading to place the students in term of their reading abilities, we did correlation analysis on both the EPT reading section and the HELP reading. Table 9 demonstrates that the EPT reading and HELP reading have a rather strong correlation, r =.914, p = .00, two-tailed, which can be interpreted as that that a student scoring high on one test will very possibly score high on the other test. A scatterplot of the students' scores on both tests will present more visualized picture of this positive and neat correlation (see Figure 3).

Table 9

Correlations between the EPT Reading and the HELP reading

		EPT reading	HELP reading
EPT reading	Pearson Correlation	1	.914(**)
	Sig. (2-tailed)		.000
	N	13	13
HELP reading	Pearson Correlation	.914(**)	1
	Sig. (2-tailed)	.000	
	N	13	13

** Correlation is significant at the 0.01 level (2-tailed).

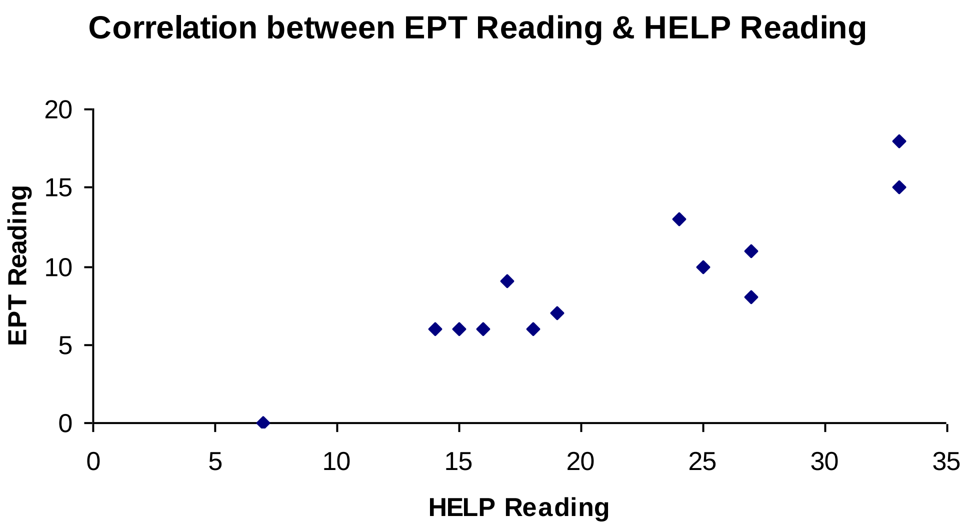


Figure 3: Scatterplot of the students' scores on the EPT reading and HELP reading tests

Therefore, based on the analysis on the HELP reading and the correlation between the HELP reading and the EPT reading, we can claim with confidence that the HELP reading is a quite reliable test to be kept for use. However, due to the relatively low item reliability index, the test might not be quite suitable to be used as placement test since the placement test is always applied to new groups of students. In addition, the strong correlation between the HELP reading and the EPT reading shows that the EPT reading can substitute the HELP reading as indicator of students' reading abilities.

HELP Writing Placement Test

In this section, we will address the rating problems which arose in the HELP writing placement test. In the interview with the Director of HELP, he expressed concerns about the lack of rating rubrics in the writing placement since the writing samples were rated based on the teachers' experience with students at each level. In order to address this issue, we re-introduced the rating rubric that HELP developed in 2006 in the first place. Instead of giving a recommended level for each writing sample in a holistic way, we designed a new scoring sheet for the raters to score in a more analytic manner. According to the level description in the rubric, the raters give a recommended level for each category. The final level given by one rater will de averaged among all eight categories. Each writing sample will be rated by two different raters.

Table 10

The Scoring Sheet for the Writing Placement Test

Category	**1**	**2**	**3**	**4**	**Level**
Fluency					
Organization					
Flow/ Cohesion					
Sentence Complexity					
Grammar/ Mechanics					
Vocabulary					
Focused Topic/ Support					
Depth of Content					

We introduced the rubric and the new rating standards in Term Spring II. 16 new students took the test. Three instructors were responsible for rating the writing samples. Each sample was rated by two different raters. One writing sample was rated by all three raters. Even though the inter-rater reliability can be rather strong, the consistency of raters' bias involved in the scoring process could influence the location of person ability estimates, and therefore to influence the placement of the students. We choose to use the Many-Facets Rasch Model (FACETS) in the analysis. Three major variables were investigated: the severity of the judge, the ability of the examinees and the difficulty of items (categories). Based on the research questions proposed in Kondo-Brown in 2007 investigating the rater bias in measuring Japanese second language writing performance, similar questions were brought up in order to guidance our analysis.

1. How variable are the overall severities among the raters? How internally self-consistent are the raters?
2. What are the relative difficulties of the 8 categories, and how consistent are the relative difficulties of these categories?
3. What are the implications for rating procedures of performance-based testing?

Table 1 provides overall information about the three variables of the examinee, the rater and the test item in terms of the Rasch measures. For the examinee, higher Rasch measures mean greater ability. For the test items, higher Rasch measure mean items are easier. For the rater, higher Rasch measures mean raters are more lenient. From table 1, we also can see the measure magnitude of the examinee ability for each scale (level). Therefore, for instance, the students measured around -4 logits will be probably recommended as level 2.

Table 11: Facets Summary of Measurement of Judge Severity, Examinee Ability and Item Difficulty

measure	+Judges	+Examinees	Scale	+ Items

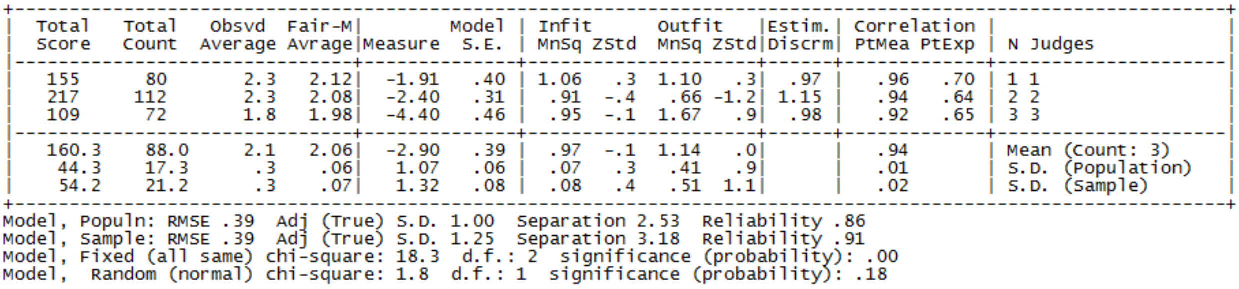
+8			(4)	
		16		
+7		1		
+6				
+5		2		
		7		
+4			(3)	
		9		
+3				
+2				
+1				F/C*, Fluency, FT/S*, G/M*, Sentence complex
		3		Vocab
0				
				Depth of Content
-1				
				Organization
-2	1			
	2			
-3				
-4		4	(2)	
	3	5		
-5		10		
-6				
		12		
-7				
-8		14		
-9		11 13 15	(1)	
		6 8		

1. How variable are the overall severities among the raters? How internally self-consistent are the raters?

Table 12 reports a detailed judge measurement summary. As stated above, the higher the Rasch measure is, the more lenient the rater is. From the negative measures, we can see that the raters are generally quite harsh on rating. Rater 1 measured -1.91 logits, is the most lenient and rater 3 measured -4.40 logits is the most hard on scoring. The difference in severity estimates between these two raters is quite large (2.49 logits). The reliability of separation index is quite high (.91), indicating a strong possibility that the raters will consistently maintain differences in overall rating severity. The Infit mean square of all three raters is measured around the expected value of 1, meaning that no rater is identified as misfitting. The fixed Chi-square calculated 18.3, with 2 df is significant, at p =.00, meaning that the raters are not equally severe in rating. In fact, the three raters do differ in severity when rating, and they consistently maintain the differences in severity when rating other similar samples.

Table 12

Judge Measurement Report



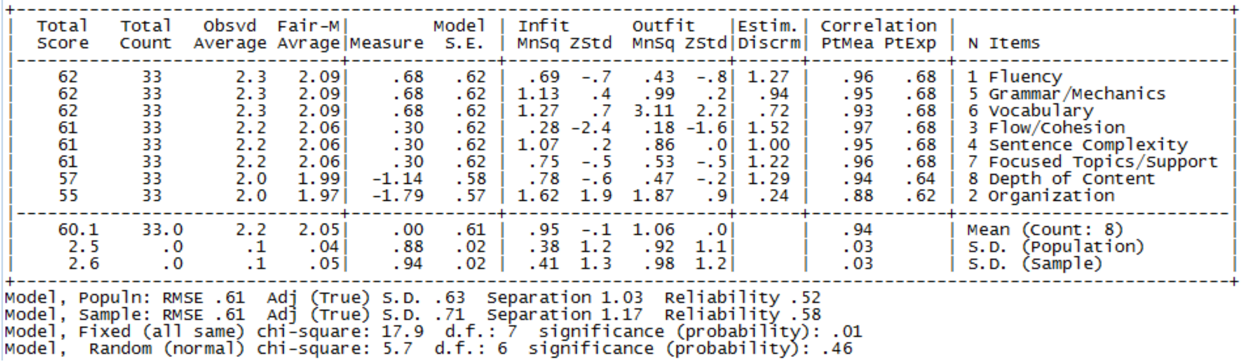
```
+---------------------------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair-M|          Model | Infit      Outfit    |Estim.| Correlation |                          |
| Score   Count  Average Avrage|Measure  S.E.  | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N Judges                 |
|------------------------------+---------------+---------------------+------+-------------+--------------------------|
|  155     80       2.3   2.12|  -1.91   .40 | 1.06    .3  1.10   .3|  .97 |   .96   .70 | 1 1                      |
|  217    112       2.3   2.08|  -2.40   .31 |  .91   -.4   .66 -1.2| 1.15 |   .94   .64 | 2 2                      |
|  109     72       1.8   1.98|  -4.40   .46 |  .95   -.1  1.67   .9|  .98 |   .92   .65 | 3 3                      |
|------------------------------+---------------+---------------------+------+-------------+--------------------------|
|  160.3   88.0     2.1   2.06|  -2.90   .39 |  .97   -.1  1.14   .0|      |   .94       | Mean (Count: 3)          |
|   44.3   17.3      .3    .06|   1.07   .06 |  .07    .3   .41   .9|      |   .01       | S.D. (Population)        |
|   54.2   21.2      .3    .07|   1.32   .08 |  .08    .4   .51  1.1|      |   .02       | S.D. (Sample)            |
+---------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .39  Adj (True) S.D. 1.00  Separation 2.53  Reliability .86
Model, Sample: RMSE .39  Adj (True) S.D. 1.25  Separation 3.18  Reliability .91
Model, Fixed (all same) chi-square: 18.3  d.f.: 2  significance (probability): .00
Model,  Random (normal) chi-square: 1.8  d.f.: 1  significance (probability): .18
```

2. What are the relative difficulties of the 8 categories, and how consistent are the relative difficulties of these categories?

Table 13 provides a detailed report of all 8 items (categories) measurement. We can see that Item 1 Fluency, Item 5 Grammar/Mechanics, Item 6 Vocabulary are scored the most leniently as they are all measured relatively higher than the rest of 5 categories, with a logit value of .68. Item 2 Organization estimated -1.79 logits is regarded as the most harshly scored category. Next to Organization, Item 8 Depth of Content is also scored quite harshly as it has a negative logit value of -1.14. In between Item 3 Flow/Cohesion, Item 4 Sentence Complexity and Item 7 Focused Topics/Support are scored relatively more leniently than Item 8 and 2, and more harshly than Item 1,5,6. From the measure range of from .68 logits to -1.79 logits, we can see that the degree of harshness is quite huge among the eight categories. The fixed Chi-square calculated 17.9 with 7 df, is significant at p = .01, demonstrating a significant variance in difficulty among the eight categories. As to the Fit value, Item 2 Organization with an Infit mean square of 1.62 is regarded as misfitting as its difficulty measure of -1.79 logits slightly exceeds the range of two standard deviations (.95 +/- [2×.41]). Despite of Item 2, other items are within the range of two standard deviations, meaning that they are scored with consistently different bias from the rater.

Table 13

Item Measurement Report



```
+--------------------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair-M|        Model | Infit      Outfit   |Estim.| Correlation |                       |
| Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtMea PtExp | N Items               |
|------------------------------+--------------+---------------------+------+-------------+-----------------------|
|   62      33       2.3   2.09|   .68    .62 |  .69  -.7   .43  -.8| 1.27 |   .96   .68 | 1 Fluency             |
|   62      33       2.3   2.09|   .68    .62 | 1.13   .4   .99   .2|  .94 |   .95   .68 | 5 Grammar/Mechanics   |
|   62      33       2.3   2.09|   .68    .62 | 1.27   .7  3.11  2.2|  .72 |   .93   .68 | 6 Vocabulary          |
|   61      33       2.2   2.06|   .30    .62 |  .28 -2.4   .18 -1.6| 1.52 |   .97   .68 | 3 Flow/Cohesion       |
|   61      33       2.2   2.06|   .30    .62 | 1.07   .2   .86   .0| 1.00 |   .95   .68 | 4 Sentence Complexity |
|   61      33       2.2   2.06|   .30    .62 |  .75  -.5   .53  -.5| 1.22 |   .96   .68 | 7 Focused Topics/Support |
|   57      33       2.0   1.99| -1.14    .58 |  .78  -.6   .47  -.2| 1.29 |   .94   .64 | 8 Depth of Content    |
|   55      33       2.0   1.97| -1.79    .57 | 1.62  1.9  1.87   .9|  .24 |   .88   .62 | 2 Organization        |
|------------------------------+--------------+---------------------+------+-------------+-----------------------|
|   60.1    33.0     2.2   2.05|   .00    .61 |  .95  -.1  1.06   .0|      |   .94       | Mean (Count: 8)       |
|    2.5      .0      .1    .04|   .88    .02 |  .38  1.2   .92  1.1|      |   .03       | S.D. (Population)     |
|    2.6      .0      .1    .05|   .94    .02 |  .41  1.3   .98  1.2|      |   .03       | S.D. (Sample)         |
+--------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .61  Adj (True) S.D. .63  Separation 1.03  Reliability .52
Model, Sample: RMSE .61  Adj (True) S.D. .71  Separation 1.17  Reliability .58
Model, Fixed (all same) chi-square: 17.9  d.f.: 7  significance (probability): .01
Model,  Random (normal) chi-square: 5.7  d.f.: 6  significance (probability): .46
```

3. What are the implications for rating procedures of performance-based testing?

From FACETS analysis of the writing test in terms of the rater severity, the difficulty of the category and the student ability, we found out that raters do differ consistently in their degree of severity, indicating that some rater scores more leniently and some rater scores more harshly. However, as the negative logit measures indicate, all raters score unanimously harshly. The magnitude of the rating severity among raters is very huge too, which reveals that the harshness of scoring among raters is quite different. However, all the raters are self-consistent in the rating severity. In respond to the different rating severity, the best solution is to have rater training. All raters need to be trained to be familiar with each category and the descriptions attached to the categories.

In terms of the category difficulty, we found out that the there is also a consistent rater bias towards different categories. Organization and Depth of Content are consistently scored more harshly than other categories and Fluency, Grammar/Mechanics and Vocabulary are scored the most leniently. Therefore, it is important to clarify each category and the description for each level in the rater training. Especially for those categories scored the most harshly and the most leniently, very detailed descriptions or specific examples will be very helpful to materialize the rating rubric.

In conclusion, the re-introduction of the rating rubric helps to establish consistent rating standards. A much more analytic rating method also helps to identify the rater bias towards each aspect of the writing samples. However, due to the existing different degree of rating severity and the rater bias towards different categories, a rater training is really important to help validate the writing placement.

Developing a speaking test

There have been measurable problems with this part of the placement results in the past, where the listening and speaking portion was identified as the least-reliable portion of the whole. The current placement test consists of a group oral interview. These interviews are usually done in groups of three or four. The assessor, usually a HELP teacher, leads a short introduction session and then approximates the test-takers appropriate level and writes a short impression or opinion of the test-takers' abilities.

This test is problematic for several reasons. The teachers used as raters are not well-trained (if they are trained at all) in eliciting speech in a group setting, about what to look for, or how to rate. Because the amount of spoken (interaction) data which can be observed in a given time is limited, the raters' judgments sometimes vary greatly. Also, the impressionistic rating is questionable because many raters do not have a clear concept of the expectations for and differences between each of the four levels at HELP. The raters may also be focusing on different aspects of student production/interaction such as grammar or pronunciation and assigning rating based on criteria completely different from another rater in the same session.

In order to create a valid and reliable test, steps were taken to assure procedural consistency and inter-rater reliability. First, we reviewed relevant literature and decided that a picture task would be the best option for our placement purposes (see "Review of Speaking Tasks" – separate document). Then, test specifications were written and agreed upon (see Appendix A). Next, a rating scale that corresponds with proficiency expectations for each level was created. The distinctions between the four curricular levels are a key part of the rater training. We also received some feedback on the scale from the teachers, curriculum coordinator, and director.

After the scale was created, two appropriate pictures were chosen. Questions were made for the interview based on the scale and the picture. Several questions for each level were created so that the interlocutor can ascertain the test-taker's proficiency after giving several questions from the appropriate level. The newly developed tests were then piloted to reveal any weaknesses or to clarify any ambiguities in the question/guideline/scale design. We received feedback from the pilot and incorporated it into the revision of the test. One important aspect of the test is that the raters must be trained thoroughly. We recommend that training take place for all teachers at HELP, whether they usually give the placement test or not. It will be important for administrators to create a consistent plan for training and re-training, and that one or two people be responsible for training. I also recommend that several samples of interviews and student performance be recorded on video for training purposes. Once the test is in full-use, a cycle of evaluation will be important to assuring that it continues to discriminate. In the first and second placement periods, I recommend that the test be compared with data from the other placement tests, assuming that these are good measures of student proficiency.

Evaluation of test design, procedures, and logistics

Written test administration and scoring procedures

As we were imputing the data for analysis, the first finding that became apparent was that there were many mistakes in the administration and scoring procedures. Some of the problems we encountered in the Michigan EPT scoring are:

1. Student A, the person scoring counted correctly and wrote the correct number right in the margin, which appears to be standard in most cases. When the person who transferred the score from the data sheet to the "whole student population" file, they interpreted the number correct as number incorrect and recorded the students' total score as 62 rather than the correct 38. There seems to be no consistency in whether the scorer writes the number correct or the number incorrect in the margin.
2. Student B, left many blanks in the answer sheet. They were counted as right. The actual score should be 33. The reported score was 51. The scorer did not count blanks as incorrect.
3. Student C, the scorer did not count four blank test items. Real score, 40, reported score is 44.
4. Student D, the scorer added the correct item number incorrectly. Real score, 35, reported score, 45.

As this is only a partial accounting of the mistakes that we found in scoring and addition, we estimate that well over 10% of students were scored incorrectly. This finding called for a drastic revision of scoring and administration procedures. We began by designing the first draft of test administrating and scoring procedures. In the following placement exam, Spring II 2009, we piloted the new materials. We received feedback and revisions from the pilot and revised the new procedures. The new procedures can be seen in Appendix B. We recommended to the administration that all faculty be trained on the administration and scoring of the new placement test. This training took place for all HELP staff on Friday May 8th, 2009. Subsequently, we recommend that new teachers be trained in the first placement administration after they are hired.

Procedures and logistics

We observed one administration of the placement test and found that there were some procedural issues that arose in the administration. As many of the documents and procedures were new, some staff did not yet know what to do. In order to compensate for these issues, we created a few new documents and procedures. First, we made a checklist for everything that should be included in the test boxes for testing day so that both the staff person that prepares the box and the person who administers the test can be sure that they have all the materials they need in advance. Second, we organized the test folders to assure that outdated materials were not cluttering the current test materials. Finally, we made a student-placement data coversheet that includes all the students' placement scores and the final placement decision that is placed in each students' file after they are placed. All three of these documents were created to facilitate placement preparation and information retention at HELP.

Recommendations & Conclusions

- The HELP Reading test can be replaced by the Michigan EPT.
- The HELP Writing test can be shortened to 30 minutes.
- The HELP Speaking test should be reviewed within the next six months to ensure that it is functioning well.
- Scorers of the Writing and Speaking tests should continue to use the rubrics for scoring.
- New teachers should be trained to administer and score the tests properly.

- One person should be in charge of testing at HELP.
- One person should be in charge of the testing on placement day. That person should be able to answer any questions and solve most problems.

References

Bond, T. G. & Fox, C. M. (2007). *Applying The Rasch Model: Fundamental Measurement in the Human Science*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Johnson, N. (2006). *Test Evaluation and Development: A Case of Language Placement Exams. Unpublished*. University of Hawaii, Hawaii, HI

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19 (1) p. 3-31.

Appendix A

General Specifications for the HELP Speaking Test

OVERALL DESIGN	
Purpose	To create a measure of foreign language proficiency in speaking skills corresponding to HELP's four course levels (100-400). To assess the candidates' ability to communicate their ideas, express opinions, and demonstrate grammar, pronunciation, and vocabulary skills in the areas of language for study, work and social life.
Intended population	Candidates who take the HELP placement examinations Age: 18+ Sex: both male and female Non-native speakers of the target language
Intended decisions/Stakes	To make statements regarding the candidates' ability to speak at levels 100-400 for placement decisions.
Response format	Guided Interview and Picture Response Task
Number of Examiners	2 (each acts as both interlocutor and rater)
Number of Candidates	Option 1 - one Option 2 – two
Number of tasks	1
Rating Scale type	Analytic scale; the criteria are described at each level in the realms of • Fluency and coherence • Communicative interaction • Grammar • Pronunciation • Vocabulary
Reporting type	Scores translated into level placement recommendation

TASK TYPE: INTERVIEW	
Format	Guided Interview and Picture Response Task
TASK DEMANDS	
Purpose	To evaluate candidates' ability to respond to questions, communicate their ideas, express and justify opinions
Response format	Live one-to-one candidate to interlocutor interaction
Known criteria	Candidates are informed of the procedure in advance
Time constraints	6 minutes per candidate; no time for preparation
Intended operations	Demonstration of structural (grammar, vocabulary, organization), linguistic and discourse knowledge, the use of complementary strategies

ADMINISTRATION: Option 2 (Monologue & Candidate to Candidate Interaction)	
Physical conditions	• In the classroom • The candidate and interlocutor/ rater are seated across from each other. • FOR THE INTERVIEW, the picture and prompt questions are face down on the examiner's desk. FOR THE PICTURE RESPONSE the picture is placed face-up on the examinee's desk. The prompt questions are face-up in the interlocutor's desk. • Candidates enter room and are introduced to the interlocutor. • The candidate is seated at the desk facing the examiner as they interact with each other. • The interlocutor introduces themselves and begins with a few questions to become familiar with the candidate. (3 minutes) • The interlocutor places the picture in front of the candidate and asks a general open-response question: "Please describe the picture for me." Or "What is happening in this picture?" • The interlocutor ascertains a level, and asks several follow up questions from the prompt questions, or questions appropriate to the dialogue that was opened by the candidate. • The interlocutor will ask questions from at least three levels to gauge the candidate's response. (5 minutes)

	• The interlocutor will end the interview and escort the candidate to the door, then return to their desk and score the candidate in the 5 categories. • When the interlocutor is ready, another candidate will be escorted into the room.

SCORING	
Scoring plan	Each category from the criteria guidelines is awarded a separate score, which is averaged to make a level recommendation for the speaking skill.
Criteria / rating scale	Analytic scale; the criteria are described at each level in the realms of • fluency and coherence • communicative competence • grammar • pronunciation • vocabulary
Rater Selection	Qualifications: must be a HELP instructor or staff Experience: must have read the administration and scoring procedures and have been trained
Rater training & standardisation	Assessors should be trained prior to any test event.
Rating procedures	Scores are marked on the scoring sheet. Assessors work independently of each other.
Rating conditions	All rating should be done ‘live’ at test event.
Moderation	A set percentage of all tests should be monitored or recorded to ensure reliability.
Statistical analysis	The reliability of the rating procedure will be estimated using both correlations and assessor agreement statistics (percentage agreement).

Appendix B

HELP Placement Test Administration Procedures
Michigan ELI English Placement Test

The English Placement Test (EPT) is a 100-item multiple-choice test. It contains problems testing listening comprehension of short utterances (using a recorded CD), grammar in conversational contexts, selection of high frequency vocabulary to fit the context of single sentences, and reading comprehension of sentences. Administration time is approximately 75 minutes.

The EPT is used to group students into homogeneous ability levels as they enter an intensive English course. Three parallel forms of the EPT can be used for pre-testing and post-testing. HELP began using the EPT in Fall 2008 for more efficient placement of incoming students into one of HELP's four levels.

SUPPLIES: Enough testing booklets and answer sheets for the number of examinees
Recorded CD and CD player for Listening section
Scoring templates for form used

1. **Seating Examinees**
   - Count the number of examinees taking the test. Make sure everyone has been admitted and seated.
   - Close the door of the testing room at the time the test is scheduled to begin.

2. **Distributing answer sheets and test books**
   - First, hand an EPT answer sheet to each examinee. Make sure each examinee clearly writes down his/her first and last name, the test date, and the form of the test (A, B, C).
   - Second, hand an EPT test booklet to each examinee. Make sure every student gets only one test booklet. They are not to open the test booklet until you say to do so.

3. **Reading directions to examinees**
   - Have examinees open the test booklet to the first page. Read aloud the general directions below. Be sure you read at a volume and pace that allows examinees to clearly understand the instructions. Do not deviate from these directions or answer any questions while reading.

   [Read aloud]
   This examination is designed to measure your mastery of the English language. There are four different kinds of problems: listening comprehension, grammar, vocabulary, and reading comprehension. There are 100 problems: 20 listening, 30

grammar, 30 vocabulary, and 20 reading. The question and answer choices are in this test booklet, but you should mark your answers on the separate answer sheet. DO NOT MAKE ANY MARKS IN YOUR TEST BOOKLET.
Now you will begin with the listening comprehension problems.

- At this point, turn on the CD player for the appropriate Form of the test that is being administered—A, B, or C.
- Following the listening section, read the instructions for ***Grammar, Vocabulary and Reading Comprehension*** and example **III, IV** and **V** on Page 3 of the test booklet to the examinees. Answer any questions the examinees have.
- Announce that the examinees may begin the Grammar section with item #21 and proceed at their own rate through the remainder of the test, and that they will have **50** minutes to complete the test.
- Write the beginning and ending of the testing time clearly on the blackboard.

4. Questions during the test

- Examinees are not allowed to talk or leave the testing room during the test. If an examinee has questions about the test, he/she should raise their hand to notify the test supervisor. No questions concerning the content of the test may be answered.
- Examinees should be notified **10** minutes prior to the test end by the test supervisor.

5. Collecting testing materials

- At the end of the testing session, collect all the answer sheets first and then all the test books from each examinee individually.
- Students should remain seated until all the answer sheets and test books are collected.

6. Dismissal

- Before examinees are permitted to leave the testing room, notify examines of the time and place of the next test or activity on the agenda.

Appendix C

HELP Placement Test Administration Procedures

Writing Placement Test

SUPPLIES: Enough testing booklets and 3 blank notebook sheets for the number of examinees

7. Seating Examinees

- Count the number of examinees taking the test. Make sure everyone has been admitted and seated.
- Close the door of the testing room at the time the test is scheduled to begin.

8. Distributing answer sheets and test books

- First, hand 3 <u>blank response sheets</u> to each examinee.
- Second, hand <u>directions sheet</u> to each examinee. Make sure every student gets **only** one.
- Read aloud item #1 on the directions sheet. Make sure everyone writes down their names and the test date clearly.

9. Reading directions to examinees

- Read aloud items #2 &#3 on the directions sheet. Be sure you read at a volume and pace that allows examinees to clearly understand the instructions. Do not deviate from these directions or answer any questions while reading.

- Announce the testing time clearly and write it down on the blackboard. The *Writing Placement Test* will take **30** minutes to finish. The time for reading directions should not be counted into the real test-taking time.
- After reading the directions, check examinees' understanding of the directions before the test starts. Answer questions only about the directions.

10.Questions during the test

- Examinees are not allowed to talk or leave the testing room during the test. If one has questions about the test, he/she should raise their hand to notify the test supervisor. No questions concerning the content of the test may be answered.
- Examinees should be notified **5 minutes** prior to the test end by the test supervisor.

11.Collecting testing materials

- At the end of the testing session, collect all the response sheets first and then all the direction sheets from each examinee individually.
- Students should remain seated until all the response sheets and directions sheets are collected.

12.Dismissal

- Before examinees are permitted to leave the testing room, notify examinees of the time and place of the next test or activity on the agenda.

Appendix D

HELP Placement Test Scoring Instructions

English Placement Test

NOTE: Instructors in charge of scoring English Placement Test should follow these instructions:

1. Use a colored pen provided by the office for scoring.
2. Use “**X**” to mark on the left side of the items answered incorrectly. Be sure you **only** mark **the items with wrong answers**. Make your marks clear and noticeable.
3. Be cautious of **the items without any answers.** Mark them before you cover the answer sheet with the plastic master key.
4. Please give a score to each individual section. There are four sections. Items 1-20 are Listening, Items 21-50 are Grammar, Items 51-80 are Vocabulary, and Items 81-100 are Reading.
5. Add the scores up and put the total on the upper right-hand side of the answer sheet. Circle the total score. Make it clear and noticeable.
6. Re-count the score of each section and the total score. Make sure there are no mistakes.
7. Transfer the scores for each section to the report sheet for that student.

Appendix E

HELP Placement Test Scoring Instructions

Writing Placement Test

Directions:

1. Read through and become familiar with the scoring guidelines and rubric before reading any essays.

2. Pick up a student essay and a scoring sheet; write the student's name and your name on the scoring sheet.

3. Read the essay completely. Give a level recommendation (100-400) in each of the 8 categories represented on the scoring guidelines.

5. When you are finished scoring, add up the scores and divide by number of categories. Write the average at the bottom of the page. Round to the nearest whole number and give a level recommendation.

6. Staple your scoring sheet to the back of the essay.

7. Each essay should have at least two raters. If there is a discrepancy in the level recommendation, a third reader should rate the essay.

Example:

Writing Placement Test – Scoring Sheet

Student name: _______________________________

Rater name: ______________________________

Category	**100**	**200**	**300**	**400**	**Level**
Fluency					
Organization					
Flow/ Cohesion					
Sentence Complexity					
Grammar/ Mechanics					
Vocabulary					
Focused Topic/ Support					
Depth of Content					

Total______/8

Average Level _______

Level recommendation _______