Ryota Iwasaki

Soo Joung Youn

SLS 490

December 12, 2012

The Final Paper

For my final paper, I decided to test the content validity of the HELP placement test. I chose this topic because I found the most interest in validity. I chose HELP placement test the target content because I work at HELP and HELP placement test has been used for many years without any revision or evaluation.

HELP stands for Hawaii English Language Program. HELP is part of UH system. HELP is similar to NICE program, but focuses on academic preparation. HELP is open to everybody who has diploma from high school. Most of the students who attends HELP aim to get into UH undergrad/grad or other schools in America.

HELP Placement test has 100 multiple questions. There are 4 different categories: Listening, Grammar, Vocabulary, and Reading. There are 20 listening questions, 30 grammar questions, 30 vocabulary questions, and 20 reading questions. HELP placement test is Norm-reference test. Purpose of this test is to place new students to their recommended level, which range from 100 to 400. Stakeholders’ age varies from 18 to 30+. The nationalities vary from Asia (Japan, China, Korea, Taiwan) to Europe (Italy, Sweden) and other countries (Brazil, etc.) The English education backgrounds do not matter when they take test.

I collected data from over 300 students and it is from year 2010 to present. Since the test has time limit, I saw behaviors that indicated several students could not finish test on time. The behaviors were 1. Answers were totally blank after certain question. 2. Serial incorrect answers in a row for more than 10 questions towards the end of the test. These behaviors would influence the evaluation of the data, so I would omit. Since Reading section was at the end of the test, so there was more of this kind of behaviors than other sections.

For this research, I evaluated each category, but I mostly paid attention to Grammar because Grammar was my interest and I chose grammar was most important feature of language when I took the diagnostic test at the beginning of the class. Grammar section had very interesting result. Here is the histogram of score/frequency:

The mean of Grammar was 21.2117, the median was 22, the mode was 22, and the midpoint was 26. The chart shows that the distribution is negatively skewed. There is more students getting higher score than the mean, so the test was little bit too easy. Here is the table of the IFs and ID of this section:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | IF | IF UPPER | IF LOWER | ID |
| I21 | 0.73 | 0.94 | 0.49 | 0.45 |
| I22 | 0.90 | 0.98 | 0.76 | 0.22 |
| I23 | 0.85 | 0.96 | 0.67 | 0.29 |
| I24 | 0.86 | 0.97 | 0.75 | 0.22 |
| I25 | 0.67 | 0.92 | 0.43 | 0.49 |
| I26 | 0.60 | 0.84 | 0.31 | 0.53 |
| I27 | 0.64 | 0.83 | 0.43 | 0.40 |
| I28 | 0.52 | 0.75 | 0.30 | 0.44 |
| I29 | 0.78 | 0.95 | 0.54 | 0.41 |
| I30 | 0.77 | 0.94 | 0.61 | 0.33 |
| I31 | 0.82 | 0.94 | 0.72 | 0.23 |
| I32 | 0.74 | 0.94 | 0.49 | 0.45 |
| I33 | 0.83 | 0.98 | 0.62 | 0.36 |
| I34 | 0.71 | 0.93 | 0.39 | 0.54 |
| I35 | 0.75 | 0.99 | 0.41 | 0.58 |
| I36 | 0.73 | 0.92 | 0.49 | 0.43 |
| I37 | 0.61 | 0.90 | 0.21 | 0.70 |
| I38 | 0.66 | 0.95 | 0.34 | 0.61 |
| I39 | 0.70 | 0.91 | 0.44 | 0.47 |
| I40 | 0.74 | 0.81 | 0.60 | 0.22 |
| I41 | 0.71 | 0.96 | 0.45 | 0.51 |
| I42 | 0.71 | 0.86 | 0.47 | 0.39 |
| I43 | 0.63 | 0.85 | 0.44 | 0.41 |
| I44 | 0.64 | 0.87 | 0.38 | 0.49 |
| I45 | 0.86 | 0.96 | 0.71 | 0.25 |
| I46 | 0.61 | 0.85 | 0.31 | 0.54 |
| I47 | 0.47 | 0.72 | 0.19 | 0.53 |
| I48 | 0.82 | 0.96 | 0.63 | 0.33 |
| I49 | 0.46 | 0.75 | 0.21 | 0.55 |
| I50 | 0.71 | 0.93 | 0.49 | 0.44 |

Most of the questions have IF of above .60. Many questions are around .80-.90 so the questions are also easy. The highlighted questions had the lowest IDs, which indicates the questions did not fulfill its purpose: To distinguish Higher-proficiency students and lower-proficiency students. I would throw away highlighted questions. I think factors that led to this behavior is that Asian English education systems, such as Japan, China, and Korea, focus and concentrate on grammar when teachers introduce English at school. Grammar was the 2nd section, so there is no worry for not finishing on time, so there is few or no distraction. I looked at the questions and questions were not well constructed, the distractors were very obvious and I could find slight hints from the question. I think this section needs to be revised due to the distribution and the IFs/ID of the test items.

Next is the Listening section. Here is the chart:

The mean of this section was 12, the mode was 11, the median was 12, and the midpoint was 20. Since this test has only 20 questions, the mean seems to be good amount, because it is close to the 50%. The distribution seems to have normal distribution / negatively skewed, but much closer to the normal distribution, comparison to grammar section. There is more students scoring middle range than high range. There is one student who got 0 correct. This would throw the data, but without this student, the 4-Ms would be higher. Here is the IFs and ID:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | IF | IF UPPER | IF LOWER | ID |
| I1 | 0.79 | 0.92 | 0.64 | 0.28 |
| I2 | 0.79 | 0.91 | 0.60 | 0.31 |
| I3 | 0.60 | 0.82 | 0.37 | 0.45 |
| I4 | 0.62 | 0.79 | 0.40 | 0.39 |
| I5 | 0.53 | 0.75 | 0.33 | 0.41 |
| I6 | 0.53 | 0.73 | 0.29 | 0.43 |
| I7 | 0.54 | 0.67 | 0.37 | 0.29 |
| I8 | 0.63 | 0.94 | 0.30 | 0.64 |
| I9 | 0.52 | 0.82 | 0.26 | 0.56 |
| I10 | 0.59 | 0.82 | 0.31 | 0.51 |
| I11 | 0.57 | 0.87 | 0.29 | 0.58 |
| I12 | 0.74 | 0.86 | 0.62 | 0.25 |
| I13 | 0.53 | 0.84 | 0.22 | 0.63 |
| I14 | 0.48 | 0.70 | 0.26 | 0.43 |
| I15 | 0.50 | 0.65 | 0.40 | 0.25 |
| I16 | 0.43 | 0.66 | 0.28 | 0.37 |
| I17 | 0.79 | 0.96 | 0.57 | 0.39 |
| I18 | 0.66 | 0.88 | 0.46 | 0.42 |
| I19 | 0.62 | 0.82 | 0.41 | 0.41 |
| I20 | 0.67 | 0.87 | 0.50 | 0.37 |

IFs and ID are healthier than grammar section. There is several low IDs, but the IF-UPPER and IF-LOWER has good gap, indicating questions did good job distinguishing students. I think this section was hard, but “good” hard. Listening is one of the essential skills to language communication, so this test did a good job on figuring out which students needs more classes to prepare them for academic success. I think this section also needs little revision, but compared to grammar section, revision would be minor. I would throw highlighted questions away due to low ID.

Next section is Vocabulary. On the next page, there is the chart:

The mean of this section was 19.97, the mode was 25, the median was 21, and the midpoint was 26. The mean was “good” because it was near 50% (there is 30 questions), but the mode was high. Distribution is also negatively skewed. The test might have been little too easy. Here is the IFs and ID:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | IF | IF UPPER | IF LOWER | ID |
| I51 | 0.91 | 0.98 | 0.82 | 0.16 |
| I52 | 0.69 | 0.89 | 0.46 | 0.43 |
| I53 | 0.84 | 0.96 | 0.64 | 0.32 |
| I54 | 0.69 | 0.99 | 0.36 | 0.63 |
| I55 | 0.79 | 0.94 | 0.57 | 0.37 |
| I56 | 0.62 | 0.92 | 0.28 | 0.64 |
| I57 | 0.86 | 0.99 | 0.66 | 0.33 |
| I58 | 0.66 | 0.85 | 0.48 | 0.37 |
| I59 | 0.85 | 1.00 | 0.61 | 0.39 |
| I60 | 0.56 | 0.86 | 0.32 | 0.54 |
| I61 | 0.44 | 0.63 | 0.27 | 0.35 |
| I62 | 0.48 | 0.81 | 0.13 | 0.69 |
| I63 | 0.85 | 0.93 | 0.75 | 0.19 |
| I64 | 0.63 | 0.88 | 0.32 | 0.56 |
| I65 | 0.84 | 0.99 | 0.58 | 0.41 |
| I66 | 0.63 | 0.84 | 0.43 | 0.41 |
| I67 | 0.72 | 0.98 | 0.38 | 0.60 |
| I68 | 0.57 | 0.75 | 0.34 | 0.40 |
| I69 | 0.61 | 0.76 | 0.38 | 0.38 |
| I70 | 0.82 | 0.95 | 0.62 | 0.33 |
| I71 | 0.49 | 0.79 | 0.18 | 0.62 |
| I72 | 0.45 | 0.74 | 0.20 | 0.54 |
| I73 | 0.72 | 0.93 | 0.49 | 0.44 |
| I74 | 0.79 | 0.98 | 0.53 | 0.45 |
| I75 | 0.35 | 0.49 | 0.22 | 0.27 |
| I76 | 0.72 | 0.96 | 0.34 | 0.62 |
| I77 | 0.67 | 0.96 | 0.30 | 0.66 |
| I78 | 0.42 | 0.67 | 0.18 | 0.49 |
| I79 | 0.68 | 0.87 | 0.39 | 0.48 |
| I80 | 0.66 | 0.87 | 0.40 | 0.47 |

There are questions with extremely low IDs (Item 51 and 63). These questions need to be thrown away because it is not doing its purpose. Other than that, rest of IDs seems to be OK. There are several questions that have slight high IFs, but IDs are acceptable.

However, the distribution tells that this test is negatively skewed and little bit easy. This section also needs to be seen by professionals.

Last section is Reading. Here is the chart:

The actual mean was 11, but the mean include data who did not finish on time, so without those students, the mean was 12, the mode and the median was 12, and the midpoint was 19. The distribution is little bit normal distribution and negatively skewed, but better distribution than Grammar. This section is hard to evaluate because there were many students who did not finish on time. Here is the IFs and ID:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | IF | IF UPPER | IF LOWER | ID |
| I81 | 0.76 | 0.92 | 0.61 | 0.31 |
| I82 | 0.76 | 0.97 | 0.42 | 0.55 |
| I83 | 0.74 | 0.92 | 0.53 | 0.39 |
| I84 | 0.62 | 0.91 | 0.27 | 0.64 |
| I85 | 0.47 | 0.74 | 0.22 | 0.52 |
| I86 | 0.65 | 0.88 | 0.42 | 0.46 |
| I87 | 0.62 | 0.91 | 0.42 | 0.49 |
| I88 | 0.64 | 0.87 | 0.34 | 0.53 |
| I89 | 0.52 | 0.72 | 0.31 | 0.40 |
| I90 | 0.62 | 0.87 | 0.34 | 0.53 |
| I91 | 0.45 | 0.69 | 0.24 | 0.45 |
| I92 | 0.52 | 0.81 | 0.27 | 0.54 |
| I93 | 0.64 | 0.91 | 0.31 | 0.60 |
| I94 | 0.66 | 0.88 | 0.40 | 0.48 |
| I95 | 0.60 | 0.88 | 0.25 | 0.63 |
| I96 | 0.60 | 0.90 | 0.24 | 0.67 |
| I97 | 0.46 | 0.77 | 0.18 | 0.60 |
| I98 | 0.59 | 0.88 | 0.27 | 0.61 |
| I99 | 0.53 | 0.82 | 0.21 | 0.62 |
| I100 | 0.48 | 0.72 | 0.20 | 0.52 |

All the test questions are above .20, which is uncommon because other sections had several questions with ID of .2~. This is because students did not finish, so there are fewer item correct for low-proficiency group. Once again, there are un-intended factors to this section, but it is hard to determine if this is good section or not, but this section tells that several students could not finish the test on time. Test should have enough time for test takers to finish because not enough time is the un-intended negative error for test takers.

Overall, the test has enough validity, measuring what it intends to measure and test items are fulfilling the purpose of this test. However, I say this test needs to be revised because several questions had too low IDs and low IDs cannot fulfill the purpose of NRT. The distributions show most of the tests are negatively skewed. Also, the test was too long for certain students. The reliability between sections was .9 and it is very high amount, but with a lot of test takers, it was not surprise to get this amount for reliability. The fact that this test hasn't been revised for long time, I suggest having professionals evaluate this test and get this test revised.