A Valid and Reliable English Proficiency Exam: A Model from a University Language Program in Taiwan

James M. Sims

Abstract

Assessing language proficiency is not only an important task but also a sensitive issue. This is especially true in Taiwan and much of Asia where testing is strongly ingrained in the culture. There are several commercially produced proficiency exams available on the market today. But, these exams are costly, only offered at limited times, and may not be appropriate for the needs of some language programs. As a result, many universities are in the process of creating their own language proficiency exams. However, there are few models for educational institutions to follow when creating their own proficiency exams. This paper presents the steps a university in Taiwan followed to create an English proficiency exam with a high reliability, appropriate validity and strong correlation to the Test of English as a Foreign Language (TOEFL). This paper includes the six procedures used for developing the language exam: (1) determining the purpose of the test, (2) designing test specifications, (3) constructing test items, (4) evaluating and revising test items, (5) specifying scoring procedures, and (6) performing validity (content, construct, and concurrent) and reliability (split-half and Cronbach’s alpha) studies. Finally, the paper concludes with a discussion of the changes to test specifications to better reflex changes in the English ability of current university students in Taiwan. It is hoped that this paper will serve as a model for other schools that want to create their own language proficiency exams.

Keywords: language proficiency test; language program; language test construction; placement exam
1. Introduction

Assessing language proficiency is not only an important task but also a sensitive issue. This is especially true in Taiwan and much of Asia where testing is strongly ingrained in the culture (Cheng, 2005; Choi, 2008; Pan, 2013; Qi, 2007; Shohamy, 2001; Watanabe, 2004). According to Pan (2014), in an attempt to improve students’ English aptitude and their global market competitiveness, many Asian counties now require their college and university students to reach a certain level or score on English proficiency exams in order to graduate. Not to be left behind and with the encouragement of the Ministry of Education, most universities in Taiwan now require their students to pass certain language proficiency tests such as the General English Proficiency Test (GEPT) or the Test of English for International Communication (TOEIC) in order to graduate (Pan, 2013; Roever & Pan, 2008; Tsai & Tsou, 2009). As a means to help students who did not achieve the required threshold on these exams, universities are providing “support/alternative/complementary measures” (Pan, 2014) to help students to fulfill their language exit requirements. These alternative methods include both remedial classes and internal in-house proficiency exams.

In addition to playing a gate-keeping role, general proficiency exams are used as placement tests with the purpose of placing students into appropriate levels or sections of a language curriculum or program (Alderson, Clapham, & Wall, 1995; Brown, 2004; Hughes, 2003). Somewhat more controversial, proficiency exams are used in a pre-post format to assess improvements in students’ general language ability as a means to evaluate the effectiveness of certain curriculums or language programs (Brown, 2004).

Proficiency exams are used for numerous purposes and there are several commercially produced proficiency exams available are the market today. But, these exams are costly, only offered at limited times, and may not be appropriate for the needs of some programs. As a result, many universities are in the process of creating their own language proficiency exams. However, there are few models for educational institutions to follow when creating their own proficiency exams.
The purpose of this paper is to present the procedures a university followed to create a language proficiency exam with an appropriate validity, high reliability, and strong correlations to established standardized exams. First, the paper outlines the procedures that were followed to create the three sections (grammar, reading, and listening) of the exam. Next, the steps that were used to determine validity and estimate reliability are presented. Finally, the paper concludes with a discussion and explanation of the changes to test specifications to better assess the current language ability of university students in Taiwan.

2. Literature Review

There is no clear definition or agreement on the nature of language proficiency. Many researchers (Bachman & Palmer, 1996) prefer the term “ability” to “proficiency” because the term “ability” is more consistent with the current understanding that specific components of language need to be assessed separately (Brown, 2004, p. 71). However, there is general agreement that both terms are made up of various related constructs that can be specified and measured. This paper, like Bachman and Palmer (1996), endorses the notion of language ability which consists of separate components embodied in four skills: listening, speaking, reading, and writing.

McNamara (2000) suggests integrating several isolated components with skill performance as a means to demonstrate the more integrative nature of language ability. Hence the proficiency test presented in this paper was constructed around language components (grammar) and skill performances (reading and listening). Likewise, it was designed to “measure general ability or skills, as opposed to an achievement test that measures the extent of learning of specific material presented in a particular course, textbook, or program of instruction” (Henning, 1987, p. 196).

In the creation of this paper, the author reviewed the following publications: Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments (Educational Testing Service, 2013); the International Test
Commission’s International Guidelines for Test Use (International Test Commission, 2000); and the International Language Testing Association’s Guidelines for Practice (International Language Testing Association, 2007). Many of the recommendations from these documents were incorporated into the model presented in this paper. These include the critical stages in the planning and development of an assessment of English proficiency for individuals who have learned English in a foreign-language context as well as the development and scoring of selected- and constructed-response test items, analyzing score results, and conducting validity research. These are presented in the next sections of the paper.

As recommended by Brown (2004), the following six procedures for developing a language test were employed in the construction of the exam: (1) determine the purpose of the test, (2) design test specifications, (3) construct test items, (4) evaluate and revise test items, (5) specify scoring procedures, and (6) perform validity and reliability studies.

3. Exam Construction

3.1 Determine the purpose

The language proficiency exam was created to serve three purposes. The first purpose was to place students into different levels of Freshman English for Non-majors (FENM) classes based on their language ability and to determine which students could qualify to waive FENM. The second purpose was to create a diagnostic tool to help identify students’ weaknesses and strengths. As Brown (2004) pointed out, besides offering information beyond simply designating a course level, a well-designed placement test may also serve diagnostic purposes. The third purpose of the exam was to evaluate the effectiveness of the FENM program by using it in a pre and post test format to measure improvements in students’ general language ability after one school-year of instruction. The program’s coordinating committee and the department council decided that a general English proficiency exam with three constructs (grammar, reading, and listening) could accomplish these purposes.
In order to accomplish these purposes, two time factors had to be considered. The biggest factor was that the results of the exams taken by nearly 3,600 students needed to be calculated in a very short period of time. The turn-around time between students taking the exam and their first class can be less than two days. Secondly, only a 70-minutes period during the freshman orientation was allotted for the exam.

With these two time factors in mind, it was decided to create a multiple-choice exam composed of 60 questions. Each question had four plausible choices, but only one correct answer. A multiple-choice format was selected because scores could be calculated quickly via the use of computer cards for answer sheets. As Bailey (1998) stated, multiple-choice tests are fast, easy, economical to score, and can be scored objectively. To reflect its major purpose, the exam was named the New English Placement Exam (NEPE).

3.2 Design test specifications

The NEPE is constructed to assess three constructs: Grammar, Reading, and Listening. The Grammar Section (20%) is composed of two cloze paragraphs with 10 questions each for a total of 20 points. The Reading Section (40%) is composed of two short passages with 5 questions per passage and one longer passage with 10 questions for a total of 40 points. The Listening Section (40%) is composed of three parts: Short Dialogues (7 questions), Short Passages (7 questions), and Appropriate Response (6 questions).

The following guidelines were used in the construction the multiple-choice items: (1) each item measured a specific objective; (2) both the question and distractors were stated simply and directly; (3) the intended answer was the only correct answer; and (4) the answer and distracters were lexically and grammatically correct, were in a parallel grammatical structure (i.e., either pairs of complete sentences or pairs of phrasal forms), and were in pairs of equal lengths with no choice being significantly longer or shorter than the others.
3.2.1 Grammar section

Hughes (2003) stresses the importance of a grammar component in a placement exam to place students in the appropriate class level. In addition, the knowledge of students’ grammatical ability is useful for determining what gaps exist in their grammatical repertoire. The grammar section of the NEPE was designed to measure students’ ability to recognize language that is appropriate for standard written English. With this in mind, the grammar section of the NEPE focused on proper verb tense, subject-verb agreement, adjectives of comparison, count versus non-count nouns, object pronouns, possessive pronouns, relative clauses, conjunctions, and passive voice. These grammar points were judged sufficient to give adequate separation between grammar scores so that students could be placed into appropriate class levels and areas of weakness could be identified.

The two cloze passages with multiple-choice answers, as well as the words, phrases and grammar points to be tested were not selected randomly, but were based on linguistic criteria as suggested by Chapelle and Abraham (1990). The topics of the cloze passages were of a general nature considered to be known to all Taiwanese high school students so that no particular group would have an advantage. Each passage was approximately 200 words in length with one cloze blank in every sentence (about ten words apart). The chosen passages did not require a deep understanding of the content of the passage so that the questions could focus on the desired grammar item being tested. Distracters were presented in simple parallel formats or were different forms of the same words or verb tenses.

3.2.2 Reading section

The Reading Section was composed of two short passages of about 240-300 words and one longer passage of 560 words (see Appendix A for a model). The first passage and questions were less difficult than the second passage and questions, while the third passage and questions were the most difficult. This was done in order to create a distribution of reading scores that would separate the more proficient reader from the lower ones. The passages had: (1) a clear, straightforward, and factual introduction,
a simple style, and a very clear, explicit thesis statement at the end of the introductory paragraph; (2) a body with unified, coherent paragraphs headed by clear topic sentences; and (3) a clear conclusion in the last paragraph. The reading passages were expository and referential in nature, like a magazine or textbook text and somewhat academic in content (i.e., not conversational or filled with slang or idiomatic English). Moreover, the reading texts were factual, informative, and descriptive, as suggested by Alderson (2000). Each paragraph was indented and numbered, with tested vocabulary underlined and bolded.

Based on Alderson (2000), the construct of “reading ability” is considered to be made up of several skills, which can be assessed by both macro-skill questions and micro-skill questions (Brown, 2004; Hughes, 2003). The term “macro questions” refers to items designed to test students’ general understanding of a passage or paragraph, while the term “micro questions” refers to items designed to test students’ understanding of specific words and sentences. The reading passages of the NEPE had the following macro-skill questions: main idea of article, main idea of paragraphs, and inference; and the following micro-skill questions: general comprehension/details, and vocabulary in context.

The following test specifications from the program’s FENM teachers’ handbook were used for the construction of the different types of reading questions:

**Main Idea Questions:** Passages for the NEPE were organized clearly enough so that a skilled reader could get the main idea by reading the introductory paragraph, the topic sentences of the body paragraphs, and the conclusion. The main idea questions asked the reader to identify what a paragraph/passage is about in general. They accurately described the overall purpose of most of the sentences in a paragraph/passage. The correct answers for main idea questions were not too general. They did not include ideas that were beyond the topic of the paragraph/passage. Likewise, the correct answers for main idea questions were not too specific. Moreover, they were not limited to particular details, reasons, or examples given in the paragraph/passage.
Comprehension/Details Questions: These questions involved identifying a specific detail in the passage. Details were facts that were clearly stated in a passage. To answer this type of question, readers had to locate a fact and choose an answer that was a paraphrase of the appropriate fact from the passage. The paraphrase provided the same meaning but differed somewhat in vocabulary and grammar.

Vocabulary-in-Context Questions: These questions asked students to find the synonym that made the most sense when it was substituted for the word or phrase in question. Vocabulary questions did not merely test whether a student could identify a synonym or definition of the given word; students had some contextual help in choosing the correct answer. These context clues appeared both inside and outside of the sentence or paragraph in which the word appeared.

Reading for Inferences Questions: An inference was a conclusion that could be made from the details in the passage. The inference was not directly stated in the passage, but it was suggested by one or more facts or was understood as being implicitly suggested or required by the explicit text.

3.2.3 Listening section

Brindley (1998) suggested that a listening test is more likely to provide a balanced assessment if a variety of different tasks are used. Hence, the listening component of the NEPE was composed of three sections: short dialogues, short passages, and appropriate response (see Appendix B for examples). The purpose of the short dialogues and short passages was to test general comprehension of concise listening texts. The goal of the appropriate response section was to test students’ immediate listening skills through the use of an appropriate response within the context of what students heard. Based on Buck (2001), these three sections consisted of questions to assess students’ listening ability to: (1) process realistic spoken language automatically and in real time; (2) understand the main idea of the passage; (3) understand explicit information in the passage; and (4) draw inferences from a passage.
The following test specifications from the program's FENM teachers' handbook and the listening committee guidelines were used for the different types of listening tasks:

**Text and the Time:** The entire listening section, including questions, was designed to be between twenty-one and twenty-three minutes long with the recording being no longer than twenty-three minutes. The topics of the texts used in the listening section demonstrated a wide variety of subject matters, but were considered to be familiar to the students. The language of the texts was colloquial instead of formal language for writing. To avoid confusion, direct speech and the use of too many names were discouraged.

**Dialogues and Short Passages:** All dialogues (approximately 100 words) were carried out between a male and female speaker; the number of turn-taking between the male and female speakers were limited to 6 to 8 exchanges in each dialogue. All short passages had a beginning, middle, and ending, with no flashbacks; the length of each short passage was about approximately 200 words and was recorded by only one person. Each dialogue had one WH-question, while each short passage had two. The questions were content-based instead of grammar-or vocabulary-based and tested comprehension of the material heard; that is, they did not allow anyone to get the right answers simply by calling on logic or general knowledge, or by knowing the meaning of a specific word. The questions tested students’ general understanding rather than their memories. There were some questions that ask for specific information or recall as well as those that asked for global understanding. The questions for dialogues were in present tense, while the questions for passages were in past tense. The names of characters were not mentioned in the questions. Instead, “the man” and “the woman” or “the mother” and “the son” were used.

**Appropriate Response:** The appropriate response section was a dialogue between a man and a woman. During the course of the conversation, one person did not know what to say. Students had to choose the most appropriate response from the choices given. The choice had to make sense in terms of what was said previously in the dialogue. Only one of the speakers asked “What should I say?” There were 3-5 lines of dialogue
or exchanges between each appropriate response question; that is, the questions were evenly spaced throughout the dialogue. The choices were of equal length and brief, keeping in mind that students only had 5 to 7 seconds to read all four choices. The questions only tested information that had been heard. In other words, as the dialogue progressed from beginning to end, the questions tested students on the moment in the dialogue previous to the question being asked.

3.3 Construct test items

The FENM program at this university puts a lot of time and energy into the creation of their midterm and final exams and over the years it has accumulated a rich test bank of materials. These exams are designed to measure students’ general reading and listening proficiency levels and are not progress or achievement tests based on specific classroom materials and instructions. All of the items for the NEPE were selected from this test bank and were not novel items.

The items in the exam bank had gone through a rigorous review process. First, individual teachers developed items using the above specifications as a blueprint. Next, test committees composed of five to seven experienced Freshman English teachers reviewed and revised each item. Then, the test item was submitted to a coordinating committee of three teachers who were not directly involved in the production of the exam item to ensure it was valid based on a comparison of test specifications and the test item. Finally, after the exams the test committees evaluated and revised or discarded items based on item analysis.

3.4 Evaluate and revise test items

Since items had been used on FENM midterm or final exams in previous years, grammar cloze texts, reading passages, scripts, and questions were evaluated then accepted based on item difficulty, item discrimination, and distractor analysis from these previous administrations. Item difficulty was used to ensure a near normal distribution of scores, to make each subsection of the exam progressively more difficult, and to create an overall total mean score of between 55-60%. This meant that some items were easy
while others were difficult. But, the total mean of all items (the total exam) would range between 55-60%. For a 100 point exam, this would ensure near normal distributions of scores. Item discrimination was used to differentiate appropriately between high and low test takers. Distractor analysis was used to make sure that distractors were efficiently distributed. Each distractor was plausible, but incorrect, and had some respondents. For ideal items, 55-60% of the students selected the correct answer (item difficulty of 0.55 to 0.60) and each distractor was selected by 10-25% of the test takers. For all items, students in the lower group (bottom 27% of all test takers) than in the upper group (top 27% of all test takers) selected distractors. In other words, every distractor was chosen by a higher percentage of low achievers than high achievers. This was done to ensure that the higher level students were not being drawn to an incorrect answer by poor distractors.

Even though all of the items had been used before, they were retested with 223 students from eight different sections of FENM (2 high, 4 mid, and 2 low) in the spring semester before the exam was first used. This ratio of 2 high, 4 mid, and 2 low sections reflected the makeup of the total FENM population (Sims, 2008). Item analysis, from this pilot study indicated that one question needed to be re-examined and revised. For this question, 8.2% of the students chose A, 52.9% chose B, 4.9% chose C, and 34.3% chose D. The correct answer was D, but most students chose B. It was discovered that the word “not” had been omitted from the question. The original question was, “According to the passage, which of the following statement is TRUE?” It was revised to “According to the passage, which of the following statement is NOT TRUE?” which was the intended question.

3.5 Specify scoring procedures

The Grammar Section (20%) was composed of 20 questions for a total of 20 points. The Reading Section (40%) was composed of 20 multiple-choice questions for a total of 40 points. The Listening Section (40%) was composed of three parts: Short Dialogues (7 questions), Short Passages (7 questions), and Appropriate Response (6 questions) for a total of 40 points. Computer cards would be scanned on a SCANMARK 2000 and scores would be calculated by using an EXCEL program.
3.6 Perform validity and reliability studies of the NEPE

As students’ scores on the NEPE served as a significant indication of their English proficiency level and determined which classes they should attend or if they qualified to waive FENM, there is a great need to offer empirical evidence to support the valid uses of the test scores. What follows are the validation approaches that were used to estimate the validity and reliability of the NEPE.

3.6.1 Validity

Validity is a complex concept, yet it is a major issue in validating a language exam. Validity in general terms refers to how appropriately a test measures what it is supposed to measure. In order to determine whether the NEPE was an appropriate instrument, three methods were used to investigate the validity of the test. First, a content validity study was conducted to examine all test items on the NEPE. Second, a construct validity study by means of an exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) was performed after the first administration of the NEPE to investigate clustering among the observed variables from the test performance. Third, a cross-comparison correlation study between the NEPE and other established standardized exams was conducted to investigate the concurrent validity of the NEPE.

**Content Validity:** A content validity study was conducted based on a comparison of test specifications and test content. The test specifications, or skills meant to be covered, were presented above. Following Hughes’ (2003) recommendations, these comparisons were made by three Freshman English teachers who were trained in language teaching and testing, but were not directly involved in the production of the exam. These teachers concluded that the exam items were appropriate measures of the desired test specifications for grammar, reading, and listening.

**Construct Validity:** In addition to investigating content validity, factor analysis was used to investigate the construct validity of the NEPE. The factor analysis consisted of a three-step process. First, EFA was performed
for the purpose of determining the best factor structure for the NEPE. Next, the best solution from EFA was tested with CFA. Finally, factorial invariance was assessed by using Comparative Fit Index (CFI), Goodness of Fit Index (GFI), and the root square error approximation (RMSEA) to ascertain any deviations of the derived model.

As Bachman (2004) and Shin (2005) stated, an EFA is a statistical procedure used to investigate clustering or patterns of commonality among the observed variables. In the NEPE, each section was developed according to the test specifications. Thus, exam items were classified into different variables based on what they were intended to measure. The intended design of the NEPE was to assess three constructs: grammar (G), reading (R), and listening (L) with the sub-structure, or tasks, of the NEPE being further divided into two grammar factors composed of two cloze paragraphs (G1 and G2), three reading factors composed of three passages (R1, R2, and R3) and three listening factors composed of three parts: short dialogues (LSD), short passages (LSP), and appropriate response (LAR). Table 1 presents more detailed information about the variables from the NEPE.

Standard EFA procedures were as follows. In the preliminary step, a matrix of product-moment correlations among the variables was devised. Then principle components analysis was used to extract the initial factors. The scree plot and eigenvalues obtained from the initial extractions were examined as an indication of the number of factors represented by the variables.

<table>
<thead>
<tr>
<th>Table 1: The variables from the NEPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>1. Grammar cloze 1 (G1)</td>
</tr>
<tr>
<td>2. Grammar cloze 2 (G2)</td>
</tr>
<tr>
<td>3. Reading passage 1 (R1)</td>
</tr>
<tr>
<td>4. Reading passage 2 (R2)</td>
</tr>
<tr>
<td>5. Reading passage 3 (R3)</td>
</tr>
<tr>
<td>6. Short dialogues (LSD)</td>
</tr>
<tr>
<td>7. Short passages (LSP)</td>
</tr>
<tr>
<td>8. Appropriate response (LAR)</td>
</tr>
</tbody>
</table>
data. After that, principle axes were used for extraction with the number of factors equal to one above and one below the number of factors indicated by the elbow of the scree plot. These extractions were rotated to both orthogonal and oblique solutions. The final step was to determine the optimum number of factors to extract from simple structure and meaningful interpretation.

Based on the results of the EFA, three factors were extracted, and the three-factor solution was used to meet the goals of interpretability and was preferable in terms of comprehensibility. The three factors were characterized according to the factor loading patterns. Factor 1 was a grammar factor because the two grammar variables loaded heavily on the first factor. Factor 2 was a reading factor, which had high loadings from the three reading variables. Factor 3 was a listening factor because the three listening variables loaded heavily on the third factor. The EFA results for the NEPE are presented in Table 2.

CFA was conducted to determine (1) the loadings on each path leading from a construct to an observed variable (task), (2) the total variance of each observed variable explained by the corresponding latent variable (the $r^2$), and (3) the model fit indicators, specifically GFI, CFI, and RMSEA. The three-factor model derived from EFA and path loadings from CFA are shown in Figure 1. The correlations between the factors and the observed variables are in Table 2.

### Table 2: The EFA results for the NEPE

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar cloze 1 (G1)</td>
<td>0.63</td>
<td>0.31</td>
<td>0.32</td>
</tr>
<tr>
<td>Grammar cloze 2 (G2)</td>
<td><strong>0.88</strong></td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>Reading passage 1 (R1)</td>
<td>0.05</td>
<td><strong>0.85</strong></td>
<td>0.19</td>
</tr>
<tr>
<td>Reading passage 2 (R2)</td>
<td>0.39</td>
<td><strong>0.64</strong></td>
<td>0.14</td>
</tr>
<tr>
<td>Reading passage 3 (R3)</td>
<td>0.41</td>
<td><strong>0.52</strong></td>
<td>0.40</td>
</tr>
<tr>
<td>Short dialogues (LSD)</td>
<td>0.20</td>
<td>0.28</td>
<td><strong>0.75</strong></td>
</tr>
<tr>
<td>Short passages (LSP)</td>
<td>0.24</td>
<td>0.30</td>
<td><strong>0.71</strong></td>
</tr>
<tr>
<td>Appropriate response (LAR)</td>
<td>0.13</td>
<td>0.04</td>
<td><strong>0.84</strong></td>
</tr>
</tbody>
</table>

*Note. High loadings for each variable are in bold and underlined.*
variables (path loadings) range between 0.56 and 0.76. The path loading for G1 (0.72) was higher than for G2 (0.60). The reading path loadings (R1: 0.56; R2: 0.61; R3: 0.75) increased from the first task to the third. The listening tasks generally had the highest path loadings (LSD: 0.75; LSP: 0.76; LA: 0.63). The GFI (0.996) and CFI (0.995) indicators were above the threshold of .90 while the RMSEA (0.026) was below the threshold of 0.05 (Byrne, 1994). The results of these three fit indices indicate that the three-factor structure model was a good fit for the dataset.

Both EFA and CFA confirmed that the NEPE measured three constructs: grammar, reading, and listening. Moreover, the factor structure indicated that constructs were measured by the corresponding variables (tasks) within the NEPE: the two grammar clozes measured grammar; the three reading passages measured reading, and the three listening tasks
measured listening. This solution supported the test structure proposed by the test designers, and served to validate the inferences made based on scores on this exam. Quite simply, the test structure derived via factor analysis reflected that intended by the designers.

**Concurrent Validity.** Brown (2004) suggests that statistical correlation with other similar independent exams is another widely accepted form of evidence for a test’s validity. The results of such a cross-comparison correlation analysis of 66 freshmen who took both the NEPE and the TOEIC indicated that the total scores of the NEPE had a high correlation \((r = 0.89)\) with the total scores of the TOEIC. The listening section \((r = 0.83)\) and grammar/reading sections \((r = 0.83)\) also had strong correlations. A similar analysis of 601 freshmen who took both the NEPE and the Intermediate GEPT indicated that the total scores of the NEPE also had a positive correlation \((r = 0.76)\) with the total scores of the GEPT and moderate correlations for listening \((r = 0.70)\) and reading \((r = 0.61)\). It should be noted that for the correlation studies, the NEPE grammar and reading scores were combined because both the TOEIC and GEPT assessed grammar on their reading sections and do not have an independent grammar component. In short, the correlation results imply that the NEPE has strong concurrent validity similar to that of the TOEIC and a strong to moderate concurrent validity to the Intermediate GEPT.

### 3.6.2 Reliability

Reliability is the extent to which a test is consistent in measuring whatever it does measure. In other words, if a student were to take the same exam on two different occasions, the results should be similar. A split-half method was used to estimate the content reliability of the NEPE, while a Cronbach’s alpha approach was used investigate the item variance reliability (see Discussion and Conclusion). For the split-half method, the exam was divided into two equivalent halves with each half composed of matching content, or skills. For example, each test item was carefully matched with a similar type of question from the other half. Questions that dealt with the main idea of paragraphs were paired up with other questions designed to measure the understanding of the main ideas of paragraphs. The same was
A Valid and Reliable English Proficiency Exam: A Model from a University Language Program in Taiwan

done for comprehension/details questions, vocabulary in context questions, and inference questions. As for the grammar section, similar grammatical points were paired together. A similar procedure was followed for the listening section. The Spearman-Brown split-half reliability coefficient was calculated to be \( r = 0.873 \), while Cronbach’s alpha reliability coefficient was \( r = 0.868 \). According to Hughes (2003), the NEPE can be considered a reliable instrument based on these two high reliability coefficients.

4. Discussion and Conclusion

Results from EFA supported a three-factor solution and CFA confirmed the three-factor model as the best fit for the data. This model reflected the test structure posited by the test designers (grammar, reading, and listening sections) and provided evidence for the construct validity of the exam. The results of factor analysis also found that the three sections of the NEPE reflected constructs that were factorially distinct. While the separability of listening from reading and grammar is widely accepted (Bae & Bachman, 1998; Hale, Rock, & Jirele, 1989; Shin, 2005; Song, 2008), the distinctness of grammar from reading is more controversial. For example, Tomblin and Zhang (2006) found the two constructs were distinct, whereas Römhold (2008) found grammar and reading grouping together. An explanation for the separation of reading from grammar found in this study may be found in an examination of the content of NEPE. The reading items focused on main ideas, specific details, and vocabulary in context, and did not require a deep understanding of the syntax or grammar. On the other hand, the grammar items mostly dealt with appropriate verb tense, subject-verb agreement, count versus non-count nouns, possessive pronouns, conjunctions, and passive voice, which did not require a deep understanding of the content of the passage.

Both a split-half method and an item variances approach were used to estimate the internal reliability of the NEPE. First, since the NEPE was designed to measure different abilities (i.e., grammar, reading, and listening) and different aspects (i.e., grammar points and reading and listening skills) of the same abilities, it was reasonable to estimate the
internal consistency with a Spearman-Brown split-half reliability coefficient (Bachman, 2004). Second, since the NEPE was designed for the scores of items to be independent and parallel measures with similar variances, a Cronbach's alpha method was also appropriate (Bachman, 2004). Quite simply, the split-half method estimated the reliability based on the content of the exam, while the Cronbach's alpha reliability coefficient estimated the reliability based on the variance of the individual items. As reported in the results section, both the reliability based on both the “content” \( r = 0.873 \) and the “variance” \( r = 0.868 \) of the NEPE were high.

The FENM program at this university had been using placement exams composed of the same three constructs (grammar, reading, and listening) and same test specifications for many years. However, the program found that the test specifications were no longer appropriate. First, test specifications needed to be changed to better reflect the English ability of current university freshmen in Taiwan. Sims and Liu (2013) and Sims (2012) found that the English listening ability of incoming university freshmen in Taiwan have improved significantly over the last two decades, while students’ grammar and reading ability have declined. Second, the test specifications of the old placement exams, no longer provided an appropriate distribution of reading scores for the lower half of the test takers and the top half of the listening scores. In other words, reading and listening scores were becoming skewed. To use an ABCD grading scale analogy, the old placement exam could separate the “A” from “B” readers from “C” and “D” readers, but could no longer separate “C” readers from “D” readers. Similarly, the old listening section could separate “C” listening scores from “D” listening scores, but could no longer separate the “A” listening scores from the “B” listening scores. Because of these two reasons, the program decided to revamp the test specifications for the reading and listening sections to the ones presented in this paper.

On the old placements exams, the reading section used to be composed of two long passages with ten questions each. It was decided to replace one of the long passages with two shorter passages with five questions each. Under the new test specifications, the first passage and questions would be composed of items with a difficulty level averaging
higher than .60, meaning that on average over 60% of the test takers would answer the questions to the first passage correctly. Basically, one long passage and questions was turned into two shorter passages with the first passage and questions being significantly easier. This was done to separate “C” readers from “D” readers and to account for the current decline in incoming university students’ reading ability in Taiwan.

The old listening sections used to be composed of three components: one long story, one long dialogue, and the appropriate response task. The purpose of the long story and dialogue was to test general comprehension of extended listening texts. On the old placement exams, both of these listening tasks used to be played twice. Now the one long dialogue has been changed into 7 short dialogues while the one long story has been replaced with three shorter passages. The shorter listening text meant that students’ memory was no longer a factor and that all listening items could be heard only once. This in turn would make the listening section more difficult and also better separate the “A” listening scores from the “B” listening scores.

As indicated by the item analysis (see Appendix C) from the first administration of the NEPE, the item difficulty, item discrimination, and distractor analysis, or item variance, were all suitable. For example, the tasks of each section of the exam were progressively more difficult. The first cloze passage and questions (0.56) of the grammar section was easier than the second cloze (0.47); while the first reading passage and questions (0.62) were less difficult than the second passage and questions (0.55) and the third passage and questions were the most difficult (0.49); the listening section was likewise progressively more difficult (short dialogues 0.68, short passages 0.64, and appropriate response 0.62). The overall mean score (57.8%) of the NEPE fell within the desired range. Item discrimination and distractor analysis were all determined to be appropriate based on item analysis methods suggested by Hughes (2003).

The proficiency exam presented in this paper for the most part is a criterion-referenced test because it was principally designed to assess the language components and skills presented previously. To a less extent, it a norm-referenced tests because it was designed to have near normal distribution and a continuum of scores. This was needed in order to divide
3,600 students into 120 sections of FENM. The NEPE was able to do both because it had clearly defined criteria to measure and the results of previous item analysis.

For a test to be appropriate and effective, it needs to be practical. According to Brown (2004), a practical exam: (1) is not too expensive, (2) remains within appropriate time constraints, (3) is relatively easy to administer, and (4) has a scoring/evaluation procedure that is specific and time-efficient (p. 19). The NEPE exam met all these criteria.

5. Implications and Limitations

The guidelines and procedures presented in this paper incorporated many of the recommendations of the publications reported in the literature review for the creation of a “general” proficiency exam with the primary purpose of a placement exam. However, they can also serve as a model for the creation of other proficiency exams. For example, over the course of a year, the language center where this study took place is asked by the university to produce numerous “general” English exams: winter transfer exam, summer transfer exam, MA entrance exam, extension entrance exams, staff promotion exams, etc. Each year after these exams have been used, they are posted online. As a result, exams cannot be reused and new exams have to be created each year. By adjusting the difficulty level of the test content, the program found the procedures outlined in this paper to be an effective model in the creation of these exams.

The NEPE has been used for two school-years for the three purposes for which it was designed. Informal feedback from teachers indicated that the exam placed students into appropriate levels of FENM. Future research is needed to confirm this. On a diagnostic level, exam results indicated that students have problems with inference questions on the reading section and with certain grammar points. As a result, teachers have focused more on helping their students to understand and make inferences and initial data from the post-test indicate that these efforts are paying dividends. As for the grammar points, future studies are needed to better identify these weaknesses and find ways to help students with them. Also, future research
is needed to investigate the appropriateness of the NEPE for the purpose of a pre- and post-evaluation of the FENM program. But, preliminary results from the first year are favorable. These results need to be confirmed with data from the second year (the end of the current school-year). The NEPE was not designed on the criteria the Common European Framework of Reference for Languages (CEFR). Yet, it is worthy to note that in the correlation study between the NEPE and the Intermediate GEPT reported in the concurrent validity section above, 158 out of 165 students (95.8%) who scored 60 points or higher on the NEPE passed the first stage (listening and reading) of the Intermediate GEPT. According to the LTTC website, the first stage of the Intermediate GEPT corresponds to the B1 level of the CEFR. This suggests that students who score 60 and above on the NEPE may be at least at the B1 level of the CEFR for listening and reading. However, future research should be conducted to better understand the NEPE in terms of the different levels of the CEFR.

The construction of the NEPE was no minor accomplishment. The designing of test specifications required identifying points to be tested and then determining appropriate and practical means to assess these items. The construction of items was a time-consuming process that required the evaluation and revision of items. However, all this attention to details of construction resulted in a cost-effective, time-saving, accurate and reliable instrument.

References


A Valid and Reliable English Proficiency Exam: A Model from a University Language Program in Taiwan


Appendix A

Sample Reading Passage and Questions

The following example for FENM midterm and final exams from the program teachers’ handbook was used as a model in the creation of the reading section.

1. It’s not just technology that’s changed in the last couple of hundred years. Most of us expect that we’ll finish our educations and get a job, and that we may change jobs several times throughout our lives. We expect that we’ll find a mate, get married, and perhaps have children someday, and that those children will grow up and have their own lives, with their own families. Our lives are filled with change—new places to live, new jobs, new friends. Although it’s very ordinary to expect these things, it’s also true that someone who lived before the Industrial Revolution might think we were crazy for having such ideas. The Industrial Revolution played a big part in changing lives from predictable ones in rural settings to the more diverse existences we now enjoy.

2. We’re used to the idea of constant change, but throughout most of human history, this has not been the norm. For thousands of years, people’s lives were much like their parents’ had been. A vast majority of the world’s population lived in the same village or on the same farm their whole lives. Boys grew up learning their father’s work so that they could continue it. Girls’ fates were decided by their parents’ choice of a husband for them, and there was mostly only one career: wife and mother—raising the next generation who would again live in the same place, doing the same things.

3. The Industrial Revolution changed many of those patterns. One of the greatest changes was urbanization, the move to cities by large numbers of people who lived in rural areas. Beginning in the 19th Century in Great Britain, factories needed huge numbers of workers; these factories were built in cities because they needed electricity, gas, water, and roads—things that were not available in rural areas.
People came to take these jobs perhaps not knowing that doing so would change whole societies. No longer would men expect to work their fathers’ trades. No longer would women be just another piece of property to be auctioned to the man with the most money. Even though most of the work in the factories required little training, it was still training that had to be standardized, so that it could be repeated to lots of workers. This kind of training was one of the foundations of mass education, and it was available—for the first time in history—equally to men and women.

Industrialization was no picnic, though. Often the work was dangerous and hard, and there were many unscrupulous factory-owners who cheated their workers. The work was often boring and repetitive, and workers were replaceable, so they felt disconnected from their work, partly because factory-work sometimes meant that a worker assembling part of a machine would never see the finished product. Perhaps the most tragic effect of industrialization, though, is pollution. For many years after the beginning of the Industrial Revolution, pollution from factories filled many rivers; more pollution came from the many people who came to work in them. The air became clouded not only from the factories, but also from the products they produced—machines that created more pollution.

If we can dream of deciding to live in almost any city in the world, doing work that our parents cannot even imagine, we get a sense of historical perspective from recognizing that these are ideas that were unheard-of only a couple of hundred years ago, before the Industrial Revolution. From modern-day urban culture, to free public education, to equal rights for women, the effects of the growth of 19th-Century factories truly changed the world in ways the first industrialists could not have foreseen.
Sample questions, distractors, explanations, and answers for the essay above

What is the main idea of this passage?
A) There are many changes in life; people get jobs and have families.
B) The Industrial Revolution changed the way a lot of people live.
C) The major effect of the Industrial Revolution is the change in women’s status.
D) 19th-Century industrialists could not have foreseen the changes they would make.

The correct answer is “B.” “A” is too general. The passage is about some changes wrought by the Industrial Revolution, not just “many changes,” or about how people get jobs or have families. “C” and “D” are both details mentioned in the passage.

What is the main idea of paragraph 2?
A) We lead lives that are different from those lived by people before the Industrial Revolution.
B) A couple of hundred years ago, most people were farmers who never went to cities.
C) Women have much more power in society now than they did before the Industrial Revolution.
D) The Industrial Revolution was the greatest change in all of history.

The correct answer is “A.” “B” is an improper inference, and anyway the paragraph is not about people being “farmers,” and the conjecture that they never went to cities is unfounded. “C” is incorrect because it is an inference drawn from a detail in the paragraph. However good an inference it may be that women’s lives are better, it is not the main idea; the paragraph talks about men and women. “D” compares the Industrial Revolution to other changes; the paragraph says nothing about this comparison.
What is the main idea of paragraph 5?
A) Industrialization did not mean that people got to eat in parks.
B) Factory-workers during the Industrial Revolution had to work hard.
C) Air and water-pollution are effects of the Industrial Revolution.
D) The Industrial Revolution had some bad effects, too.

The main idea of this paragraph is that while the Industrial Revolution gave people new opportunities, there were also some costs to society; thus, “D” is the best answer. “B” and “C” are both true, and are both mentioned in the paragraph, but they are both details, not the main idea. “A” incorrectly assumes a literal meaning of an idiomatic phrase.

What does norm mean in paragraph 2?
A) new way B) different way C) best way D) ordinary way

The choice that comes closest to a synonym for “norm” is “ordinary way;” “D” is the best choice. Context is provided by contrasting “ideas” that we are “used to” with other (“non-ordinary”) ways of doing things. Further context clues are provided by the use of “ordinary” in paragraph 1, and by the use of “vast majority of the world’s people” in paragraph 2. This is a more difficult vocabulary-in-context question because the context clues are further removed from the word. The simplest form is probably the vocabulary word followed by a defining appositive, as in the next example.

What does urbanization mean in paragraph 3?
A) moving to cities B) working in factories
C) getting a new job D) making revolution

The correct answer, “A,” is defined in the appositive. Appositive definitions are usually the simplest, partly because the context-clues are nearest the word to be defined.
What does auctioned mean in paragraph 4?
A) married  B) engaged  C) sold  D) given

Context for the correct answer, “C” is given at the end of the sentence. More difficult vocabulary-in-context questions will place the contextual clues further from the word to be defined.

According to the passage, what can be INFERRED about Great Britain before the Industrial Revolution?
A) Public education was not available to both men and women.
B) Before the Industrial Revolution, there were no cities in Great Britain.
C) London, the largest city in Great Britain, is very polluted.
D) Factory-workers are usually unhappy people.

“A” is the only inference that can be defended by using the text of the essay as support. The statement that begins “For many years after” makes “C” a bad choice; there’s certainly an inference that London “was” polluted, but not that it “is.” Similarly, the verb-tense in “D” is unconnected to anything the essay says. “B” is unwarranted.
Appendix B

Sample Listening Questions
These examples were written, edited, and revised by FENM teachers on the Listening Committee.

Sample dialogue and question
Woman: Excuse me. Your singing is too loud. Do you mind keeping it down?
Man: Well, I'm a music major, and I need to practice my singing for a performance on Saturday.
Woman: For Pete's sake. I have to turn in my homework this afternoon. I can't do it with your loud singing in the background.
Man: When is your afternoon class?
Woman: At two.
Man: Oh well, in that case, I'll practice my singing after you go to your class.
Woman: I really appreciate your understanding.
Man: Sure!

What does the woman want?
A) some quiet time to complete her homework
B) someone to help her with her homework
C) a good idea for her homework assignment
D) a chance to help the man practice singing

Sample short passage and questions
During the winter vacation, Jack spent most of his time at home reading newspapers and talking on the phone. His mom couldn't stand it and yelled at him, complaining that he had not been considerate and helpful. His father had been out of work for three months and had been out looking
for work every day, while Jack stayed at home taking it easy. Hearing his mom's complaint, Jack felt sad because she didn't know all this time he had been reading the want ads in the newspaper and making phone calls, just because he was trying to find a part-time job to help his family out.

Question 1: Why was Jack's mother angry with him?
A) He hadn't been helpful to his family.
B) He hadn't been studying at all.
C) He hadn't been grateful to the family.
D) He hadn't been listening to her.

Question 2: What happened to Jack's father?
A) He took a break from his work.
B) He complained about his job.
C) He was seriously ill.
D) He lost his job.

Sample Appropriate Response
Susan and Bill are Freshmen students in the Department of Political Science at Tunghai University.

Susan: Hi Bill!
Bill: Hi Susan! How are your classes going?
Susan: Oh, pretty good, I suppose. The only class I'm having a problem with is Political Science. Our teacher is very strict and gives us lots of quizzes. He says this will help us to learn more, but I am not so sure. So, what did you do during Spring Break?
Bill: Some friends of mine invited me to join them in Kenting National Park for a camping trip.

Susan: What should I say? (WSIS?)
1. A) How disgusting!  B) How weird!
   C) How unusual!  D) How wonderful!
Bill: Yes, we had a great time. We talked a lot, we played games and we had a barbecue. I really had fun on the trip. How about you, Susan?

Susan: My Spring Break wasn't very good.

Bill: What happened?

Susan: WSIS?

2. A) I walked to the Tunghai main gate and waited for a long time.
   B) I went home and was told my parents were going to get a divorce.
   C) I got up this morning and ate my breakfast but it wasn't very good.
   D) I ran across the street to Seven Eleven to buy some new toothpaste.

Bill: You must feel terrible.

Susan: Yes, home life has gotten worse. I never want to go back to that place again.

Bill: Susan, don't get so upset! Don't cry! Look, maybe, er . . .

Susan: Bill, you don't understand. Ok. Just forget it ok?

Bill: Ok, ok. But what are you going to do for summer vacation if you don't go home?

Susan: WSIS?

3. A) Oh, I don't know, maybe I'll settle for a vacation in Hawaii.
   B) Oh, I'm not sure, maybe I'll go to Seven Eleven to get a newspaper.
   C) Perhaps, next weekend I will read a book about Political Science.
   D) Perhaps next year I will change my major and study Economics.

Bill: You can't be serious, Susan.

Susan: Look, Bill, I just don't know what to do.

Bill: Why don't you come and stay with my family? We have a large house in Kaoshiung. You can take my sister's room. She has gone to New Zealand to study for a year.
Susan:  WSIS?
4.  A) Bill, you are an idiot and I really hate you.
    B) Bill, your sister must be very beautiful to study overseas.
    C) Bill, that's sweet of you, but I want your house.
    D) Bill, that's very kind of you, but no. I couldn't.

Bill: Why not? You'll have a good time
Susan:  Nope. I am sure something will work out. I am a big girl, you know.
Bill:  Well, if you change your mind, let me know, ok?
Susan:  You still don’t understand, do you, Bill?
Bill:  What?

Susan: WSIS?
5.  A) I will go to New Zealand.     B) I can write a short story.
    C) I can take care of myself.     D) I will do it again.

Bill:  Well, er, like I say, I just want to help.
Susan: Thanks, Bill, but I have to go now. My Political Science teacher is giving us another quiz in a few minutes and I'd better run. Bye!
Bill: Yes, see you later.
Appendix C

The Results of Item Analysis for the NEPE

<table>
<thead>
<tr>
<th>Item</th>
<th>Item difficulty</th>
<th>Upper 27%</th>
<th>Lower 27%</th>
<th>DS indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95%</td>
<td>99%</td>
<td>89%</td>
<td>10%</td>
</tr>
<tr>
<td>2</td>
<td>64%</td>
<td>87%</td>
<td>41%</td>
<td>46%</td>
</tr>
<tr>
<td>3</td>
<td>57%</td>
<td>76%</td>
<td>40%</td>
<td>36%</td>
</tr>
<tr>
<td>4</td>
<td>64%</td>
<td>75%</td>
<td>53%</td>
<td>22%</td>
</tr>
<tr>
<td>5</td>
<td>56%</td>
<td>78%</td>
<td>35%</td>
<td>43%</td>
</tr>
<tr>
<td>6</td>
<td>44%</td>
<td>71%</td>
<td>21%</td>
<td>50%</td>
</tr>
<tr>
<td>7</td>
<td>32%</td>
<td>44%</td>
<td>25%</td>
<td>19%</td>
</tr>
<tr>
<td>8</td>
<td>54%</td>
<td>81%</td>
<td>32%</td>
<td>49%</td>
</tr>
<tr>
<td>9</td>
<td>60%</td>
<td>80%</td>
<td>41%</td>
<td>39%</td>
</tr>
<tr>
<td>10</td>
<td>34%</td>
<td>43%</td>
<td>27%</td>
<td>16%</td>
</tr>
<tr>
<td>11</td>
<td>56%</td>
<td>69%</td>
<td>39%</td>
<td>30%</td>
</tr>
<tr>
<td>12</td>
<td>56%</td>
<td>68%</td>
<td>49%</td>
<td>19%</td>
</tr>
<tr>
<td>13</td>
<td>38%</td>
<td>50%</td>
<td>24%</td>
<td>26%</td>
</tr>
<tr>
<td>14</td>
<td>36%</td>
<td>49%</td>
<td>29%</td>
<td>20%</td>
</tr>
<tr>
<td>15</td>
<td>45%</td>
<td>60%</td>
<td>33%</td>
<td>27%</td>
</tr>
<tr>
<td>16</td>
<td>45%</td>
<td>71%</td>
<td>25%</td>
<td>46%</td>
</tr>
<tr>
<td>17</td>
<td>70%</td>
<td>80%</td>
<td>55%</td>
<td>25%</td>
</tr>
<tr>
<td>18</td>
<td>46%</td>
<td>67%</td>
<td>28%</td>
<td>39%</td>
</tr>
<tr>
<td>19</td>
<td>55%</td>
<td>67%</td>
<td>46%</td>
<td>21%</td>
</tr>
<tr>
<td>20</td>
<td>23%</td>
<td>32%</td>
<td>18%</td>
<td>14%</td>
</tr>
<tr>
<td>21</td>
<td>73%</td>
<td>85%</td>
<td>61%</td>
<td>24%</td>
</tr>
<tr>
<td>22</td>
<td>35%</td>
<td>53%</td>
<td>24%</td>
<td>31%</td>
</tr>
<tr>
<td>23</td>
<td>70%</td>
<td>93%</td>
<td>44%</td>
<td>49%</td>
</tr>
<tr>
<td>24</td>
<td>70%</td>
<td>89%</td>
<td>46%</td>
<td>43%</td>
</tr>
<tr>
<td>25</td>
<td>64%</td>
<td>84%</td>
<td>45%</td>
<td>39%</td>
</tr>
<tr>
<td>26</td>
<td>69%</td>
<td>88%</td>
<td>48%</td>
<td>40%</td>
</tr>
<tr>
<td>27</td>
<td>35%</td>
<td>57%</td>
<td>21%</td>
<td>36%</td>
</tr>
<tr>
<td>28</td>
<td>55%</td>
<td>81%</td>
<td>36%</td>
<td>45%</td>
</tr>
<tr>
<td>29</td>
<td>56%</td>
<td>87%</td>
<td>27%</td>
<td>60%</td>
</tr>
<tr>
<td>30</td>
<td>62%</td>
<td>87%</td>
<td>35%</td>
<td>52%</td>
</tr>
<tr>
<td>31</td>
<td>57%</td>
<td>86%</td>
<td>32%</td>
<td>54%</td>
</tr>
<tr>
<td>32</td>
<td>69%</td>
<td>90%</td>
<td>46%</td>
<td>44%</td>
</tr>
</tbody>
</table>
A Valid and Reliable English Proficiency Exam: A Model from a University Language Program in Taiwan

<table>
<thead>
<tr>
<th>Item</th>
<th>Item difficulty</th>
<th>Upper 27%</th>
<th>Lower 27%</th>
<th>DS Indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>62%</td>
<td>80%</td>
<td>39%</td>
<td>41%</td>
</tr>
<tr>
<td>34</td>
<td>30%</td>
<td>49%</td>
<td>18%</td>
<td>31%</td>
</tr>
<tr>
<td>35</td>
<td>28%</td>
<td>48%</td>
<td>20%</td>
<td>28%</td>
</tr>
<tr>
<td>36</td>
<td>47%</td>
<td>78%</td>
<td>25%</td>
<td>53%</td>
</tr>
<tr>
<td>37</td>
<td>31%</td>
<td>61%</td>
<td>14%</td>
<td>47%</td>
</tr>
<tr>
<td>38</td>
<td>74%</td>
<td>95%</td>
<td>48%</td>
<td>47%</td>
</tr>
<tr>
<td>39</td>
<td>55%</td>
<td>87%</td>
<td>24%</td>
<td>63%</td>
</tr>
<tr>
<td>40</td>
<td>36%</td>
<td>60%</td>
<td>23%</td>
<td>37%</td>
</tr>
<tr>
<td>41</td>
<td>82%</td>
<td>95%</td>
<td>65%</td>
<td>30%</td>
</tr>
<tr>
<td>42</td>
<td>60%</td>
<td>84%</td>
<td>33%</td>
<td>51%</td>
</tr>
<tr>
<td>43</td>
<td>79%</td>
<td>95%</td>
<td>60%</td>
<td>35%</td>
</tr>
<tr>
<td>44</td>
<td>48%</td>
<td>72%</td>
<td>27%</td>
<td>45%</td>
</tr>
<tr>
<td>45</td>
<td>58%</td>
<td>88%</td>
<td>26%</td>
<td>62%</td>
</tr>
<tr>
<td>46</td>
<td>67%</td>
<td>89%</td>
<td>43%</td>
<td>46%</td>
</tr>
<tr>
<td>47</td>
<td>79%</td>
<td>97%</td>
<td>55%</td>
<td>42%</td>
</tr>
<tr>
<td>48</td>
<td>68%</td>
<td>89%</td>
<td>48%</td>
<td>41%</td>
</tr>
<tr>
<td>49</td>
<td>66%</td>
<td>93%</td>
<td>36%</td>
<td>57%</td>
</tr>
<tr>
<td>50</td>
<td>74%</td>
<td>95%</td>
<td>46%</td>
<td>49%</td>
</tr>
<tr>
<td>51</td>
<td>66%</td>
<td>95%</td>
<td>37%</td>
<td>58%</td>
</tr>
<tr>
<td>52</td>
<td>45%</td>
<td>69%</td>
<td>27%</td>
<td>42%</td>
</tr>
<tr>
<td>53</td>
<td>86%</td>
<td>98%</td>
<td>68%</td>
<td>30%</td>
</tr>
<tr>
<td>54</td>
<td>50%</td>
<td>77%</td>
<td>28%</td>
<td>49%</td>
</tr>
<tr>
<td>55</td>
<td>61%</td>
<td>84%</td>
<td>42%</td>
<td>42%</td>
</tr>
<tr>
<td>56</td>
<td>61%</td>
<td>74%</td>
<td>48%</td>
<td>26%</td>
</tr>
<tr>
<td>57</td>
<td>44%</td>
<td>65%</td>
<td>24%</td>
<td>41%</td>
</tr>
<tr>
<td>58</td>
<td>75%</td>
<td>95%</td>
<td>50%</td>
<td>45%</td>
</tr>
<tr>
<td>59</td>
<td>57%</td>
<td>83%</td>
<td>30%</td>
<td>53%</td>
</tr>
<tr>
<td>60</td>
<td>72%</td>
<td>94%</td>
<td>48%</td>
<td>46%</td>
</tr>
</tbody>
</table>

Note: DS Indices refers to the discriminability.