University College of Takestan

Available online at **http://UCTjournals.com**

Iranian Journal of Social Sciences and Humanities Research

UCT . J. Soc. Scien. Human. Resear. (UJSSHR)

128-134 (2017)

Volume 5,Issue3

ISSN 2382-9753

A Valid and Reliable Proficiency Exam for English language with Respect to University Language Program in Iran

Gholam-Reza Abbasian[1] and Elmira Hajmalek[2]

1Assistant Professor, Imam Ali University and Islamic Azad University, South Tehran Branch, Tehran, Iran
2PhD Candidate, Islamic Azad University, Kish International Branch, Iran

Original Article:
Received 1 June. 2017 Accepted 30 June. 2017 Published 25 Nov. 2017

ABSTRACT
Proficiency exams are used for numerous purposes and there is 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 Iran. 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 Iran. 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 exam, language program, test construction,

***Corresponding author: Abbasian**

Peer review under responsibility of **Iranian Journal of Social Sciences and Humanities Research**

1. Introduction

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. Research suggests that teachers spend from one quarter to one third of their professional time on assessment-related activities, without necessarily having learned the principles of sound assessment, according to Stiggins (2007).

Since the first English Proficiency Exam (EPE) was designed by Adams Sherman Hill, president of Harvard in 1874, who wanted students to exhibit balanced structure in English in both classrooms and on proficiency exams, until today, competency testing in English requires enormous amounts of reflection and planning. Not only English faculty but all other faculty who design writing projects in their classrooms and even students and campus governing bodies define what the particular community values about the English language for these tests since they must mirror classroom instruction and student learning. Due to this requirement of incorporating local values and perceptions of English into a test format, evaluators, often Writing Program Administrators, work to investigate the ways written English can be defined in order to be measured. Indeed, standardizing English for assessment purposes based on the values of diverse groups is no easy task for any institution of higher education.

Shohamy (2008: xiv) argues that theories and practices in language testing have been closely related to definitions of language proficiency. Consequently, the discrete-point testing era presented isolated test items; the integrative era meant discourse language, and the communicative era typically involved interaction and authentic texts. In the performance era, real-life tasks were used; and finally, alternative assessment recognizes the fact that language knowledge is a complex phenomenon, requiring "multiple and varied procedures to complement one another".

By referring to the change in theories of learning, Brown and Hudson (2002) state that discrete item tests, as seen in the multiple-choice format for instance, were possible as long as language learning was concerned with specific grammar and language skills. When more complex uses of language were aimed for, e.g. pragmatic and sociolinguistic competence, performance testing became more valid, e.g. test items which cause the examinee to perform in the language and show communicative ability for instance (2002:57).

Bachman and Palmer emphasize that there is no model language test (2010:6): "In any situation, there will be a number of alternatives, each with advantages and disadvantages". They also point out that if we want to develop language assessments where the use is justified; there need to be justification for multiple qualities (2010:63). "[A] language assessment should consist of language use tasks. In designing language assessments whose use we can justify, it is important to include tasks whose characteristics correspond to those of TLU [target language use] tasks".

To conclude, assessment in English today is based on communicative language competence and focuses on the use of language. The European Language Portfolio, henceforth referred to as the ELP, uses "can do-statements" as descriptors for linguistic proficiency, thereby emphasizing the action-oriented approach described in the CEFR, also acknowledging the learner as a central informant (Little, 2009). In spite of the description of language proficiency as language use both in the CEFR and the ELP, a great deal of work remains to be done to increase the engagement of learner agency in assessment, according to Little and Erickson (2015). They point out that "proficiency develops from sustained interaction between the learner's gradually developing competences and the communicative tasks whose performance requires him or her to use the target language" (2015:124).

English language proficiency (ELP) assessment is an extremely important aspect of English language learner (ELL) students' academic careers as the output of such assessment determines and influences their instruction, classification and promotion. Therefore, providing reliable and valid ELP assessments are most important in determining their academic progress. Assessments of ELP based on questionable measures may cause grave academic consequences. ELL students who are inadequately assessed may be misclassified with respect to their level of proficiency in English and may receive inappropriate instruction. They may even be misclassified as students with learning disabilities, which may greatly impact their academic career (see, e.g., Abedi, 2006b; Artiles, Rueda, Salazar, & Higareda, 2005). Furthermore, ELL students' level of English proficiency is an important criterion in determining their readiness for participating in the state content- based assessments such as reading/language arts, math, and science. Because state content-based assessments that are used for No Child Left Behind Act (NCLB; 2002) Title I accountability purposes are mainly constructed and field tested for students who are fluent in English, they may be subject to linguistic factors that could seriously undermine their validity for ELLs (Abedi, 2006a).

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 Iran.

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.

Assessment literacy is a term that advocates evidence-informed practice and for assessors i.e. teachers, to reflect on the effect of their teaching and assessment strategies. Assessment literacy relates to validity in testing and assessment (Popham, 2006:84):

[I]f a teacher mistakenly believes that validity resides in the test itself, the teacher will be inclined to defer to whatever results the "valid test" produces. Assessment -literate educators, however, understand that education tests merely provide evidence that enables people to make judgmentally based inferences about students.

According to Popham (2009:7), teachers who are genuinely assessment literate know both how to create more suitable assessments and are familiar with "a wide array of potential assessment options". However, Malone (2008:225) states that "there is no consensus on what is required or even needed for language instructors to reliably and validly develop, select, administer and interpret tests". A gap between language testing practice and the training of language instructors is acknowledged. The CEFR is mentioned as one useful tool to bridge the gap.

Shepard (2000) claims that teachers need help in learning to use assessment in new ways in order to develop students' "robust" understanding. All too often, the same test types are used, implying that mastery does not transfer to new situations since students have learnt to master classroom routines and not the underlying concepts. Assessment literate teachers consequently know how to choose and use the best method of assessment to fit the context, the students, the level and the subject. Validity, reliability, authenticity, washback, purpose, student impact and constructive alignment are identified as influential concepts for assessment literate teachers (Brown, 2004; White, 2009).

Washback does not only relate to products, as in assessment outcome, but also says something about participants and processes (Bailey, 1999; Hughes, 1994). Brown and Hudson (2002) mention that a multiple choice grammar test used to test communicative performance will have a very strong negative washback effect on a communicative curriculum. Washback is related to validity, and Messick (1996) states that there needs to be an evidential link between learning outcomes and test properties. In CLIL, as in the present study, such an evidential link may not be obvious as regards language. The intentional learning goals focus on content, which is a matter of validity in the CLIL approach and will be discussed later.

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).

3. Construction of Exam

3.1 purpose of exam

The language proficiency exam aim was 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. The third purpose of the exam was to evaluate the effectiveness of the FENM program by using it in a pre and posttest format to measure improvements in students' general language ability after one school-year of instruction.

3.2 Procedure for 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

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 Iranian 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. 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, 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:

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

The listening component of the NEPE was composed of three sections: short dialogues, short passages, and appropriate response. 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:

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 Validity of the NEPE

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 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.

Table 1: The variables from the NEPE

Variable	Item (question number)
1. Grammar cloze 1 (G1)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
2. Grammar cloze 2 (G2)	11, 12, 13, 14, 15, 16, 17, 18, 19, 20
3. Reading passage 1 (R1)	21, 22, 23, 24, 25
4. Reading passage 2 (R2)	26, 27, 28, 29, 30
5. Reading passage 3 (R3)	31, 32, 33, 34, 35, 36, 37, 38, 39, 40
6. Short dialogues (LSD)	41, 42, 43, 44, 45, 46, 47
7. Short passages (LSP)	48, 49, 50, 51, 52, 53, 54
8. Appropriate response (LAR)	55, 56, 57, 58, 59, 60

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.

Table 2: The EFA results for the NEPE

Variable	factor		
	1	2	3
1. Grammar cloze 1 (G1)	0.63	0.31	0.32
2. Grammar cloze 2 (G2)	0.88	0.10	0.15
3. Reading passage 1 (R1)	0.05	0.85	0.19
4. Reading passage 2 (R2)	0.39	0.64	0.14
5. Reading passage 3 (R3)	0.41	0.52	0.40
6. Short dialogues (LSD)	0.20	0.28	0.75
7. Short passages (LSP)	0.24	0.30	0.71
8. Appropriate response (LAR)	0.13	0.04	0.84

3.5 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. 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 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 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 factorial distinct. While the separablity 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 construct were distinct, whereas Römhild (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, and count versus non-count nouns, obsessive 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 108 EaGLE Journal 1(2), 2015 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 Cranach's alpha reliability coefficient estimated the reliability based on the variance of the individual terms. 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 reflex the English ability of current university freshmen in Iran. Sims and Liu (2013) and Sims (2012) found that the English listening ability of incoming university freshmen in Iran 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 Iran.

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.

References

Abedi, J. (2006a). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 377–398). Mahwah, NJ: Erlbaum.

Abedi, J. (2006b). Psychometric issues in the ELL assessment and special education eligibility. Teacher's College Record, 108(11), 2282–2303.

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

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and Developing Useful Language test. Oxford, England: Oxford University Press. 112 EaGLE Journal 1(2), 2015

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English two-way immersion program. Language Testing, 15, 380-414.

Bailey, K. (1998). Learning about Language Assessment. Boston, MA: Heinle & Heinle.

Brown, H. D. (2004). Language assessment: Principles and classroom practices. White Plains, NY: Pearson Education.

Buck, G. (2001). Assessing listening. Cambridge, England: Cambridge University Press.

Chapelle, C. A., and Abraham, R. G. (1990). Cloze method: What does it make? Language Testing, 7, 121-146.

Cheng, L. (2005). Changing language teaching through language testing: A washback study. Cambridge, England: Cambridge University Press.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. Language Testing, 25, 39-62.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). Confirmatory factor analysis of the TOEFL. TOEFL research report 32. Princeton, NJ: Educational Testing Service.

Henning, G. (1987). A guide to language testing. Los Angeles, CA: Newbury House.

Hughes, A. (2002). Testing for language teachers. Cambridge, England: Cambridge University Press.

International Language Testing Association. (2007). Guidelines for practice. Retrieved from http://www.iltaonline.com/images/pdfs/ilta_guidelines. pdf

McNamara, T. F. (2000). Communication and design of language tests. In H. G. Widdowson (Ed.), Language testing (pp. 13-22). Oxford, England: Oxford University Press.

Shohamy, E. (2001). The power of tests. Harlow, England: Pearson Education.

Song, M. Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. Language 114 EaGLE Journal 1(2), 2015 Testing, 25, 435-464.

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. Journal of Speech, Language, and Hearing Research, 49, 1193-1208.