Concurrent and Predictive Validity of Computer-adaptive Freshman English Test for College Freshman English in Taiwan

Cindy Chou

Department of English Language, Literature and Linguistics

200, Sec. 7, Taiwan Boulevard, Shalu Dist., Taichung City 43301, Taiwan

Tel: 8864-2632-8001 E-mail: cchou@pu.edu.tw

Massoud Moslehpour (Corresponding author)

Department of Business Administration, Asia University

500, Liufeng Rd., Wufeng, Taichung 41354, Taiwan

Tel: 8864-2332-3456 E-mail: mm@asia.edu.tw

Nguyen Thi Le Huyen

Department of Business Administration, University of Finance and Accountancy

La Ha Town, Tu Nghia District, Quang Ngai Province, Vietnam

Tel: 84-989-303679 E-mail: huyenvic.eco@gmail.com

Received: January 6, 2014 Accepted: January 16, 2014 Published: January 16, 2014

doi:10.5296/ijele.v2i1.4919 URL: http://dx.doi.org/10.5296/ijele.v2i1.4919

Abstract

Undoubtedly, it is always a continuing effort for colleges and universities to seek a reliable assessment tool for placing students into appropriate college Freshman English levels in order to maximize its instructional effectiveness. The assessment tool, however, is required of fulfilling cardinal principles of test validity, reliability, and practicality. This study was designed to determine the level of criterion-related validity in terms of concurrent and predictive validity of a computer-adaptive Freshman English Test (CAFET) employed as a Freshman English placement test at a university in central Taiwan. Two other English proficiency tests, the *National Joint College Entrance Exam on English Proficiency* (NJCEEE) and Taiwan's *General English Proficiency Test* Intermediate Level (GEPT-I1) were used as criterion variables for measuring the CAFET's concurrent validity. In addition to evaluating the CAFET's internal quality, individual and/or combined predictive power of the CAFET, NJCEEE, and GEPT-I1 for Freshman English course achievement of participants of School Year 2010-2011 was also investigated. Statistical procedures of Pearson's Product Moment correlation coefficients, one-way ANOVA, and simple and multiple regressions were performed. The results supported usability of the CAFET as an appropriate tool to place students into proficiency levels of the Freshman English program, especially when practicality is a prime consideration. Implications and suggestions for future research are presented.

Keywords: Placement Testing, Computer-adaptive Testing, Criterion-related Validity, GEPT, CAFET

1. Introduction

The increasing use of placement testing to place students in appropriate course level of Freshman English has been found in Taiwan's higher education. To ensure that a placement test accurately group university freshmen by their English abilities for ensuring instructional effectiveness, the test must first demonstrate that it accurately measure the students' English abilities. The same placement test result, in addition, is expected to be able to serve to a certain extent as a reliable and valid index of Freshman English course achievement.

In determining the effectiveness of a language assessment, Brown and Abeywickrama (2010) identified five cardinal principles: practicality, reliability, validity, authenticity, and washback. Among these five principles, validity which is defined as "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment" (Brown and Abeywickrama, 2010, p. 29) is considered the most important. As a form of validity, *criterion validity*, evidenced by measures of concurrent and predictive validity, is demonstrated by correlating a test score with other well-respected measures of the same construct (Brown, 2005). To ensure that a test is highly dependable in assessing students' English ability, *concurrent validity* must be established by validating the test result against independent criterion measures. In addition, to predict students' future achievement, *predictive validity* of the placement test must be reported by building a satisfactory correlation of the placement test result with course outcomes. As Napoli, Raymond, Coffey, and Bosco (1998) emphasized, "Only when sufficient criterion-related validity has been established, and points on the [tests'] score continuum can be reliably equated to some relevant cognitive or behavior skill(s), can the test user appropriately deploy the assessment tool to make placement and curriculum decisions" (p. 9). Therefore, it is crucial for colleges and universities to establish evidence of the criterion validity of tests used to place students by English proficiency and to predict academic outcomes in order to demonstrate that the placement result does enhance students' likelihood of success in Freshman English.

Due to its greater potential for solving practical measurement problems, computer-adaptive testing (CAT) is made possible due to the evolution of psychometric basis tests and Item Response Theory (IRT) models. Increasingly, standardized tests including the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE) apply IRT to estimate students' abilities. IRT principles are involved in CAT "in both selecting the most appropriate items for an examinee and equating scores across different subsets of items" (Embretson & Reise, 2000, p. 1). On CAT, test-takers receive items that are optimally selected for his/her performance level. Based on this testing feature, it is obvious to note that major advantages of CAT include: (a) individualization, (b) easy test administration to large groups of test-takers, (c) electronic scoring, and (d) reducing test time (Brown & Abeywickrama, 2010; Olea, Abad, Ponsoda, Barrada, & Aguado, 2011).

A useful way for enhancing effectiveness of Freshman English instruction is to focus on the learning points appropriate for a relatively homogeneous English ability. To that end, placement tests which help decide students' appropriate levels within a specific program become necessary. In addition, placement tests for Freshman English are often administered to incoming students in the same day of the orientation session, and then the test results must be yielded within a day or two before the semester begins. As a result, due to practicality considerations, the Computer-adaptive *Freshman English Test* (CAFET hereafter) considered being more efficient than traditional paper-based tests in terms of test time and cost and rapid scoring was adopted for the Freshman English placement purpose.

The CAFET, a 10-level criterion-referenced test which measures receptive skills of vocabulary, grammar, listening, and reading abilities, is a testing technique developed in Year 2006 by *Smarten Tech*. Each level of difficulty constitutes of 20 questions of five primary content categories: vocabulary embedded in listening (VLM) and reading (VRM), grammar embedded in listening (GLC) and reading (GRC), and reading comprehension (RO). At the beginning of the test, students are presented with a series of randomly selected questions of Level 5. A student is then advanced to the next upper level if s/he answers correctly 60% of the questions. On the other hand, the difficulty level is lowered if s/he fails to meet the 60% criterion. The test is terminated when a student fails the same level twice.

In light of the aforementioned advantages of CAT, the CAFET has been in use at a university in central Taiwan since School Year 2007-2008 for Freshman English program placement purpose. Students taking the CAFET are entering freshman students who had previously taken the *National Joint College Entrance Exam on English Proficiency* (NJCEEE hereafter) as part of their college admission requirements. Based on the CAFET scores, the students are placed into four levels of Freshman English (i.e., A, B, C, D, with Level A being the most advanced). Moreover, in order to evaluate the learning outcomes of the 2-semester Freshman English course, all students are required to take both the CAFET and the first stage (assessing vocabulary, idioms, usage, structure, and receptive skills of listening and reading) of Taiwan's *General English Proficiency Test* Intermediate Stage 1 (GEPT-I1 hereafter) upon completion of the course. During the two semesters of Freshman English, the students are required to participate in weekly language labs for zero-credit self-access online GEPT practice. Mid-term and final summative tests containing the weekly practice questions are administered.

1.1 Purpose of the Study

Only when sufficient evidence of criterion-related validity has been established, can the school administrators and instructors appropriately deploy the test to make accurate placement and instructional decisions. However, despite the relatively high level of CAFET's statistical reliability, its criterion-related validity has not been thoroughly examined. Therefore, purpose of this study was threefold. First, concurrent validity of the CAFET scores were analyzed by correlating the scores with those of NJCEEE and GEPT listening and reading. NJCEEE and GEPT-I1 were chosen because they share assessments of vocabulary, structure, and reading comprehension for determining English proficiency. Second, numerous higher education institutes are using NJCEEE scores to be cost effective for placing students into levels of their Freshman English programs. It is therefore important to determine if the NJCEEE scores can serve as a future placement tool in a Freshman English for Non-majors (FENM hereafter) program that has students ranging from D (beginning) to A (advanced) levels. Third, whether the CAFET can serve as reliable estimates of Freshman English course achievement and pass of GEPT-I1 was also investigated. To extend from this purpose and serving as a preliminary investigation, varying combinations of test measures were examined to determine a better prediction of Freshman English course performance. Guided by these purposes, three research questions were posed for investigation: (a) To what extent does the CAFET correlate with NJCEEE and listening and reading components of GEPT-I1? (b) Can NJCEEE serve as an effective placement tool in the FENM program based on the CAFET results? (c) To what extent does the CAFET predict the Freshman English course achievement? And by extension, how well do any of the test measures along or in combination better predict the Freshman English course achievement?

1.2 Significance of the Study

As aforementioned, Brown and Abeywickrama (2010) considered validity to be the most important principle for evaluating effectiveness of a language assessment. As such, accumulated knowledge of the criterion-related validity of computer-adaptive tests for placement purpose is necessary for providing evidences for its capacity to assist in making accurate placement and curriculum decisions for colleges and universities. Overall, information derived from this study will provide empirically established evidences for researchers, administrators, and instructors concerning the utility and interpretations of using the CAFET in guiding administrative and instructional decisions for the FENM program.

Most importantly, accurately placing students and predicting learning outcomes by appropriately administering a validated computer-adaptive test within a program will enhance efficient allocations of limited educational resources.

2. Literature Review

2.1 Computer-adaptive Testing

Since the 1970s, compute-adaptive testing (CAT) has been viewed as a practical alternative to paper-and-pencil testing. CAT for second language assessment (L2 CAT) is a technological method of assessment in which the computer selects and present test items to examinees according to his/her estimated language ability. In a CAT, in other words, each examinee takes a unique test that is tailored to his/her own ability level for eliminating time-consuming and inefficient tests. Nevertheless, while researchers (Bernhardt, 1996; McNamara, 1996) expressed their concerns about the appropriacy, fidelity, and comprehensiveness of computer-based testing for assessing particular language skills, others (Dunkel, 1999) were concerned about the degree to which construct-irrelevant variables such as computer-familiarity or computer anxiety might impact performance in negative ways. With numerous advantages of CAT for language assessment, questions relating to basic principles of assessment such as reliability and validity are yet to be addressed. In particular, since CATs have not been widely used for language assessment, it is relatively difficult to assess their validity.

Aiming to fill the gap of insufficient content validity of CATs, Olea et al. (2011) reported development and psychometric properties of a CAT on English listening called *eCAT-Listening*. After conducting measures of *eCAT-Listening* and questionnaire and statistic analyses of confirmatory factor analysis, ANOVAs, *t*-tests, and Pearson correlations, the authors found satisfactory psychometric properties of the *eCAT-Listening.* Among other results, the authors emphasized the test's accuracy and efficiency in relating the examinees' scores of *eCAT-Listening* to the English program level and in shortening testing time.

2.2 Criterion-related Validity of Language Tests

An increasing number of higher education institutes is searching for a reliable, cost-effective, and easily administered assessment tool for placing incoming students within a program and for evaluating learning outcomes in the program. Undoubtedly, empirically established validity can ensure that the influences and decisions made on the bases of test scores are meaningful, appropriate, and useful (Bachman, 2004). Bachman also contended that to ensure that an assessment tool is valid, criterion-related validity must be empirically established in addition to construct validity. Eda, Itomitsu, and Noda (2008) examined and reported reliability and construction and concurrent validity of the Japanese Skills Test (JSKIT) used for purposes of both program placement and evaluation. The findings revealed that the JSKIT, having proved to be a reliable measurement tool, is an ideal tool for placement and for prediction of speaking proficiency only for students of lower levels.

With regard to reading-skill related placement, assessment of the criterion-related validity of the *Computerized Placement Test in Reading Comprehension* (CPTR) was undertaken in order to provide information concerning utility and meaning of student scores on the CPTR for actual reading level and skills (Napoli, Raymond, Coffey, & Bosco, 1998). By correlating the CPTR scores with scores of Degrees of Reading Power Test (DRP), a criterion-referenced holistic measure, results indicated that the CPTR can be employed with a high degree of reliability and validity to identify basic reading proficiency skills corresponding to the demands of first-year college-level texts.

Often, an assessment tool is used for prediction of academic achievement. Songy (2007) reported how valid the *Michigan Test of English Language Proficiency* (MTELP) and the *Cultural Intelligence Test* (CFIT) can aid in the screening processes of selection of seminary candidates and of predicting academic success in Papua New Guinea. This study aimed to see if the two instruments possess the predictive power for students' overall grade point average (GPA). Though many researchers had cautioned about use of GPA as a dependent variable, Songy found both instruments useful in predicting GPA. Some subsets of the instruments such as vocabulary section of the MTELP, however, were more useful in distinguishing low achievement.

3. Method

3.1 Participants

The data presented in this study were collected in School Year 2010-2011 from participants of Freshman English for Non-English Majors (FENM) in a university in central Taiwan. The FENM course objectives are focused primarily on English language forms of vocabulary and grammar embedded in and enhanced by receptive skills of listening and reading. The FENM students took the CAFET as both entry and exit tests for purposes of placement and achievement evaluation. Two months prior to taking the CAFET for FENM placement, the students had previously taken the *National Joint College Entrance Exam on English Proficiency*. As exit tests, both CAFET and GEPT-I1 were administered upon completion of the FENM program. Only those students with complete datasets of all the measure scores were included in the analysis for this study. As a result, a total of 874 valid sets of students' scores of NJCEEE, CAFET and GEPT-I1 (including listening and reading components) and Freshman English course were used in this study. Table 1 summarizes the participants' academic background. As can be noticed, more than half of the participants (55.5%) were from FENM Level B. Moreover, an important part of the participants (37.1%) came from College of Management while the smallest group was from College of Foreign Languages (10.4%).

Table 1. Academic Background of Participants ($N = 874$)

Variable	*n*	%
FENM Level		
A	84	9.6
B	485	55.5
C	254	29.1
D	51	5.8

College		
Management	324	37.1
Humanities & Social Science	177	20.3
Science	165	18.9
Computing & Informatics	117	13.4
Foreign Languages	91	10.4

3.2 Instruments

To achieve the purposes of this study, four English proficiency test scores were used for analysis: (a) The Computer-adaptive *Freshman English Test* (CAFET), (b) The *National Joint College Entrance Exam on English Proficiency* (NJCEEE), and (c) listening and (d) reading tests of the intermediate *General English Proficiency Test* (GEPT). Among the four variables, CAFET is the criterion variables, while the other three the predictor variables.

First, the CAFET is a testing technique developed in Year 2006 by Smarten Tech., a 10-level computer-adaptive criterion-referenced test which measures vocabulary, grammar, listening, and reading abilities. Each level of difficulty constitute of 20 questions of three types: vocabulary and grammar embedded in listening and reading and reading comprehension. Both the internal reliability (coefficient α = .60 ~ .78) and test-retest reliability (r_{xx} = .55 ~ .86) are moderate.

Second, NJCEEE aims to assess high school students' vocabulary, grammar, reading and writing skill. Question types are mainly multiple choices, sentence translation and short essay; moreover, it focuses on evaluating students' reading and writing abilities which can be divided into five groups: (a) recognizing vocabulary, (b) cloze test (include grammar and vocabulary), (c) reading comprehension, (d) sentence translation, and (e) short essay writing (write paragraphs either based on pictures or description). The scores in January and July college entrance exams are calculated in transferred levels (0-15) and grades (0-100) respectively.

Finally, GEPT is a 5-level test of English language proficiency developed and administered by the Language Training and Testing Center (LTTC). The five levels of the GEPT are elementary, intermediate, high-intermediate, advanced, and superior. According to the LTTC, the GEPT targets English learners at all levels in Taiwan, corresponds to Taiwan's English education framework, meets the specific needs of English learners in Taiwan for self-assessment, and provides institutions or schools with a reference for evaluating the English proficiency levels of their job applicants, employees, or students. The GEPT covers the language skills of listening and reading at Stage 1, and writing and speaking at Stage 2 of each level (Table 2). Passing Stage 1 of a specific level is the prerequisite for registering for its Stage 2. As for this study, the listening and reading components of GEPT-I1 were used for analysis.

Table 2. GEPT Test Format and Structure (Intermediate – CEFR B1)

Stage	Module	Part	Task types	Number of items	Max. score	Time (mins.)
1	Listening	1	Picture Description	45	120	30 (approx.)
		2	Answering Questions			
		3	Conversations			
	Reading	1	Sentence Completion	40	120	45
		2	Cloze			
		3	Reading Comprehension			
2	Writing	1	Chinese-English Translation	2	100	40
		2	Guided Writing			
	Speaking	1	Reading Aloud	13~14	100	15 (approx.)

3.3 Data Analysis

In addition to sample means and standard deviations, Pearson’s Product Moment correlation coefficients between the CAFET and NJCEEE and listening (GEPT-L) and reading (GEPT-R) components of GEPT-I1 were calculated to determine the extent to which the CAFET measures the general English proficiency level. Second, in order to determine how the NJCEEE scores are distributed against the placement of students based on the CAFET results, a one-way ANOVA was calculated to determine if the differences of the students’ NJCEEE mean scores of the four levels are significant. Third, a simple regression analysis was computed to examine if the CAFET can predict the Freshman English course achievement. And finally, stepwise multiple regression was used to investigate the amount of variance in Freshman English course achievement that could be better explained from any of the four measures.

4. Results and Discussion

To answer Research Question 1, *to what extent does the CAFET correlate with NJCEEE and listening and reading components of GEPT-I1?* Descriptive statistics of the three test scores are first calculated (Table 3). Mean scores of the four sets of proficiency test scores, the CAFET, NJCEEE, GEPT-L, and GEPT-R, are 410.46 (*SD* = 131.97), 33.62 (*SD* = 12.84), 70.71 (*SD* = 17.88), and 69.12 (*SD* = 18.4) respectively. It can be seen that the data skewness is nearly zero, which indicates the normality of the data. Therefore, mean, median and mode scores are approximately at the midpoints for the three measures of GEPT-L (.01), GEPT-R (.049) and NJCEEE (.084). This can also be initially interpreted that the participants’ English proficiency were at the beginning and intermediate levels. The CAFET data, however, show a moderate degree of negatively skewness (-.794), indicating that its mean score of the participants is above the median (midpoint score). That is, the CAFET (-.794) test is originally easier than GEPT-L (.01), GEPT-R (.049) and NJCEEE (.084) because the participants' average score on this test is above the Median.

Table 3. Descriptive Statistics of CAFET, NJCEEE, and GEPT-I1 (*N* = 874)

Measure	*N*	*M*	*SD*	*Max.*	*Min.*	*Skewness*
CAFET	874	410.46	131.97	686	125	-.794
NJCEEE	874	33.62	12.84	79.33	1.67	.084
GEPT-L	874	70.71	17.88	117	21	.010
GEPT-R	874	69.12	18.4	117	21	.049

Note. The total possible scores of GEPT-L and GEPT-R = 120, CAFET = 1000, NJCEEE = 100.

Furthermore, Pearson Product Moment correlations were calculated to determine the relationships among the four measures. Table 4 shows correlations among the four measures; i.e., the CAFET, NJCEEE, GEPT-L, and GEPT-R. As a result of analysis, weak positive and statistically significant correlations were found between CAFET and the other three English proficiency tests, NJCEEE, GEPT-L, and GEPT-R (*r* = .385, .389, and .444 respectively, *p* ≤ .01). Moderate to moderately strong positive correlations were found between NJCEEE and GEPT-R (*r* = .642), and between NJCEEE and GEPT-L (*r* = .450).

Table 4. Pearson Correlation Coefficients for CAFET, NJCEEE, GEPT-L, & GEPT-R (*N* = 874)

Measure	CAFET	NJCEEE	GEPT-L	GEPT-R
CAFET	1	.385**	.389**	.444**
NJCEEE	.385**	1	.450**	.642**
GEPT-L	.389**	.450**	1	.478**
GEPT-R	.444**	.642**	.478**	1

***p* ≤ .01.

It was found that the students who did well on the computer-adaptive test, the CAFET, may not necessarily perform well on the other three paper-based proficiency tests. Of the three determinant variables, NJCEEE, in particular, was significantly and moderately correlated with GEPT listening (GEPT-L) (*r* = .450) and GEPT reading (GEPT-R) (*r* = .642). The results indicate that the higher the students' scores on the *College Entrance Exam on English* *proficiency*, the more likely he/she is to perform well on the reading comprehension component of the *General English Proficiency Test*. Both NJCEEE and GEPT-R, in particular, emphasize assessment of students' reading abilities by employing similar question mode of multiple-choice responses and paper-based type. On the other hand, the relatively weaker correlations of the CAFET and the other three tests may lie in test type. While the test type of the other three tests is paper-based, that of CAFET is computer-based, which may have been a primary reason why those who may perform well on the other three tests may not do as well on CAFET.

Next, to answer Research Question 2, *Can NJCEEE serve as an effective placement tool in the FENM program based on the CAFET results?* Table 5 shows sample means and standard deviations of the students' NJCEEE scores of the four FENM levels. As can be seen, students' mean scores of the four FENM levels decrease from the highest Level A (M = 43.95, SD = 13.02) to the lowest Level D (M = 22.82, SD = 12.29).

Table 5. Descriptive Statistics of the 4 FENM Levels' NJCEEE Scores (N = 874)

FENM Level	*N*	*M*	*SD*
A	84	43.78	13.0
B	485	35.26	11.50
C	254	29.26	12.20
D	51	22.98	12.35
Total	874	33.82	12.84

Results of one-way ANOVA indicated that there were statistically significant group differences among the NJCEEE mean scores of the 4 FENM levels ($F_{(.05;\ 4,\ 870)}$ = 48.371, p = .000) (Table 6). Furthermore, Scheffe's Post Hoc analysis indicated that the mean NJCEEE scores of all four FENM levels (A, B, C, D) were significantly different from one another. Another words, there were statistically significant differences among all fours groups of FENM on all tests (CAFET, NJCEEE, GEPT-L1).

Table 6. Analysis of Variance for Students' NJCEEE Scores among 4 FENM Levels (N = 874)

Source	*SS*	*df*	*MS*	*F*	*p*
Between	20583.883	3	6861.294	48.371	.000
Within	123407.302	870	141.847		
Total	143991.185	873			

Finally, to answer Research Question 3, *To what extent does the CAFET predict the Freshman English course achievement* and, by extension, *how well do any of the test measures along or in combination better predict the Freshman English course achievement*? Both simple and multiple regression analyses were performed. Table 7 presents results of a simple regression analysis of the CAFET scores on Freshman English course grades. The result indicates that, at the α = 0.05 level of significance, there was a weak association (R^2 = .022) between the CAFET and Freshman English grades (p = .000). Which means, CAFET explains very little of total variation of FENM course achievement. In other words, although the CAFET can be a predictor of the students' Freshman English grades, its predictive power was statistically weak.

Table 7. Simple Regression Analysis for CAFET's Predicting Freshman English Course Performance

Variable	*R*	*R*^{2}	*F*	*B*	*SE*	*β*	*t*	*Sig.*
Constant				72.780	1.047		69.542	.000
CAFET	.147	.022	19.388	.011	.002	.147	4.403	.000

Note. $p < .001$.

Furthermore, significant predictor variables were identified and weighted in terms of their contribution to the criterion variable, the FENM grades. Results of multiple regression analysis (Table 8) show that 17% of Freshman English performance could be explained significantly by the four predictors as a group (i.e., NJCEEE, GEPT-L, GEPT-R, and CAFET). In addition, stepwise multiple regression procedure was used to determine which of the four tests can better predict the FENM course performance. As a result, GEPT-R and NJCEEE were shown to be significant predictors of the FENM achievement (F = 88.838, p < .001). Moreover, the *β weights* indicate that the strengths of contribution of GEPT-R and NJCEEE to FENM achievement are .328 and .115 respectively. Their combined positive contribution to FENM achievement is moderately high ($β$ = .443), with GEPT-R ($β$ = .328) possessing the strongest relationship with FENM achievement. That is, among the four predictor measures, GEPT-R best predicted the students' FENM course achievement (p < .001).

Table 8. Stepwise Regression for Predicting Freshman English Course Achievement

Model	*R*	*R*^{2}	*Adjusted R*^{2}	*F*	*Sig.*	*β*
1	.402[a]	.162	.161	168.159	.000	.328
2	.412[b]	.169	.168	88.838	.000	.115

[a]Predictors: (Constant), GEPT-R

[b]Predictors: (Constant), GEPT-R, NJCEEE

5. Conclusion

Purposes of this study aimed to establish criterion-related validity in terms of concurrent and predictive validity for a computer-adaptive test, the CAFET, employed by a university in central Taiwan. By extension, whether the CAFET or other assessment measures can better predict FENM achievements were also investigated. Findings derived from the statistical results are presented below in the order of research questions.

First of all, correlation analyses were performance to establish concurrent validity of the CAFET. Significantly positive correlations, albeit weak, existed among the four measures of CAFET, NJCEEE, GEPT-L, and GEPT-R (*r* = .385 ~ .444). It was found that from the students who did well on the computer-adaptive test, the CAFET, may not necessarily perform well on the other three paper-based proficiency tests, and that NJCEEE and GEPT-I1 were significantly and moderately correlated. In particular, students who performed well on NJCEEE may more likely perform well on GEPT-R. Such finding seems to be reasonable in view that both NJCEEE and GEPT-R emphasize assessment of students' reading abilities by employing similar question mode of multiple-choice responses and paper-based type. The CAFET, on the other hand, correlated relatively weakly with the other three tests may due to its computerized test type. While the test type of the other three tests is paper-based, that of CAFET is computer-based, which may have been a primary reason why those who may perform well on the other three tests may not do as well on CAFET. Such a finding may be further supported by Dunkel (1999) that construct-irrelevant variables such as computer-familiarity or computer anxiety might impact performance in negative ways. To ensure that students are not disadvantaged by computerized tests, it is thus of critical importance to develop students' computer skills. In particular, students should be able to become fully familiar with keyboarding and mechanics of taking the CAFET including entering responses and using the word processor prior to taking a computer-adaptive test such as CAFET. Nevertheless, preliminary evidences of the CAFET's concurrent validity were established in this study by being significantly correlated with scores of NJCEEE, GEPT-L, and GEPT-R, which can be inferred that the relationships of CAFET and the other measures are unlike to happen by chance.

Second, in regard to using the NJCEEE as an effective placement tool, results of one-way ANOVA show that the mean NJCEEE scores of all four FENM levels initially placed by CAFET scores were significantly different from one another, which helps to confirm that NJCEEE can serve as an effective placement candidate in the FENM program based on the CAFET results. That is, in light of program placement, college authorities can rely, to a certain extent, on these two measures as reliable tools for accurately identifying levels of English abilities. Given that NJCEEE is taken in Taiwan for admissions by most prospective college freshmen, and that the CAFET meets criteria of individualization, easy test administration to large groups of students, electronic scoring, and reducing test time, these two practical assessment instruments are thus recommended for future use for making placement and exemption decisions for college-level Freshman English programs.

Finally, in terms of prediction of students' FENM achievement, results of a simple regression analysis determine that the CAFET was a statistically significant, albeit weak, contributor to variation in Freshman English course achievement. That is, the CAFET solely may not have contributed sufficiently to the prediction of students' Freshman English course achievement. This may be due to the fact that most Freshman English courses emphasize in reading and listening as well. Therefore, this study further tried to investigate if varying combinations of two or more predictor variables can make better prediction. It was found that GEPT-R and NJCEEE combined possessed stronger relationship with FENM achievement. Besides, among the predictor variables, GEPT-R best predicted the students' FENM course achievement. Nevertheless, other unspecified variables such as instructional techniques, teacher and student interaction modes, motivation, attitudes, and difficulty of coursework should also be taken into consideration in predicting the course achievement. After all, students' academic performance is too complex an issue to investigate.

Overall, evidences yielded from this study failed to support usability of the CAFET alone as a significant predictor of Freshman English course achievement. It is thus concluded that further calculations should consider the complexity of factors involved in determining course achievement. Perhaps using non-parametric calculations would serve as a better measure of correlations. Furthermore, for colleges and universities in EFL contexts who prioritize practicality issues of administrative details, costs, and time, the CAFET may serve as a reliable tool to place students into appropriate proficiency levels for enhancing instructional effectiveness of the Freshman English program. The same universities and colleges operating under limited educational budget and resources may consider employing NJCEEE scores for both FENM placement and prediction purposes before validity of computer-adaptive tests is improved.

6. Limitations and Suggestions for Future Studies

The present study was limited by the fact that, although criterion-related validity of a computer-adaptive English proficiency test, the CAFET, was evidenced, the data were collected from only an academic year, 2010-2011 of Freshman English. Besides, since CATs have not been widely used for language assessment, it is relatively difficult to assess their validity. Therefore, more data collected from Freshman English students of successive academic years for analyses is suggested. Such a continuing research effort would also help the moderately skewed nature of the data. Second, following the purpose of examining its predictive validity, the CAFET was used as a single predictor for and failed to possess much influence on the Freshman English course achievement. As discussed earlier, other predictors including more validated English proficiency tests, coursework outcomes, course instructors' observations, and students' self-reported motivation are thus suggested to be used in future studies for predicting Freshman English course achievement. Third, the participants' mean scores were below or slightly above the midpoints of the total possible scores of the three measures, NJCEEE, GEPT-I1, and CAFET (50, 60, and 500 respectively), which can be initially interpreted that the participants' English proficiency were at the beginning and intermediate levels. Therefore, it is necessary to collect more data from students of higher levels of English proficiency to confirm the effectiveness of the CAFET as a placement test for a wider range of College Freshman English programs. Finally, in addition to NJCEEE and GEPT-I1, more accurate measures of English proficiency such as TOEFL, IELTS, and

TOEIC can be used to serve as criterion variables for a stronger support of criterion validity of the CAFET as a Freshman English placement test.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017%2FCBO9780511667350

Bernhardt, E. (1996). If reading is reader-based, can there be a computer-adaptive test of reading? In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 1-10). New York, NY: Cambridge University Press.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.

Brown, J. D. (2005). Testing in language programs: A comprehensive guide to English language assessment. New York, NY: McGraw-Hill.

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology, 2*(2), 77-93.

Eda, S., Itomitsu, M., & Noda, M. (2008). The Japanese Skills Tests as an on-demand placement test: Validity comparisons and reliability. *Foreign Language Annals, 41*(2), 218-236. http://dx.doi.org/10.1111%2Fj.1944-9720.2008.tb03290.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

McNamara, T. (1996). Computer-adaptive testing: A view from outside. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 136-149). New York, NY: Cambridge University Press.

Napoli, A. R., Raymond, L. A., Coffey, C. A., & Bosco, D. M. (1998). The CPT reading comprehension test: A validity study. *Journal of Developmental Education, 22*(1), 8-14.

Olea, J., Abad, F. J., Ponsoda, V., Barrada, J. R., & Aguado, D. (2011). eCAT-listening: Design and psychometric properties of a computer-adaptive test on English listening. *Psicothema, 23*(4), 802-807.

Songy, D. G. (2007). Predicting success in academic achievement of major seminarians in Papua New Guinea: A comparison of cognitive test results and grade point averages. *Contemporary PNG Studies: DWU Research Journal, 7*, 59-71.

Copyright Disclaimer