The Use of Online Standardized Tests as Indicators of English Proficiency according to the Common European Framework of Reference (CEFR) at the English Program of Chonkanyanukoon School, Chon Buri, Thailand.

- **Ratima Sinlapachai**
- **Janpanit Surasin**
- **Raweewan Angkanurakbun**

Abstract *The purposes of this study were to investigate whether the Cambridge English Placement Test (CEPT) and the Oxford Online Placement Test (OOPT), can provide an accurate indication of Common European Framework of Reference (CEFR) levels, to examine the correlation between the two test scores and between the test scores and the students' achievement on English courses, and to explore the stakeholders' satisfaction towards this testing and their intention on utilizing the test results. The test results and achievement scores were collected from 336 students and analyzed using descriptive statistics and the Pearson Product Moment Correlation coefficient. The qualitative data were collected from 10 students, three teachers, and an administrator by using semi-structured interviews and analyzed by using a coding method.*

The findings indicated that the CEPT and OOPT did not provide consistent CEFR levels to the test takers; however, there was a positive relationship between the scores from both tests ($\rho = .69$ at $p < .01$). Moreover, the OOPT had a higher correlation with student achievement. The study also found that the students and teachers did not have enough information about the testing and CEFR and there were many factors which affected the scores negatively. Therefore, the Program should consider using other types of measurement to add to the current method and provide more testing related information to the stakeholders.

Keywords: *Online standardized tests, Common European Framework of Reference*

In 2013, there was an instruction regarding taking a proficiency test from the Office of Basic Education Commission, Thailand. According to the instruction (2013), this proficiency testing was administered at all World-Class Standard Schools in Thailand and each school received test IDs for either the Oxford Young Learner Placement Test, the Oxford Online Placement Test, or the Cambridge English Placement Test, depending on the region and level of the school. These tests reported the results in the form of total scores and CEFR level which are: A1 (the lowest), A2, B1, B2, C1, and C2 (the highest). Chonkanyanukoon School was given 120 test IDs for students to take the Cambridge English Placemen Test. The school distributed 60 IDs to the students in the English Program. However, the instruction also stated that schools could purchase more test IDs at the cost of 77 Baht each. The Program

perceived this as an opportunity for the students to be exposed to standardized tests at an economical cost; therefore, the Program decided to purchase test IDs for the Cambridge English Placement Test (CEPT) and the Oxford Online Placement Test (OOPT) for all students in the Program.

After all students at the Program completed both tests, the results were kept in the Program's records. Then, in 2014, the Ministry of Education announced a new policy to use the Common European Framework of Reference (CEFR) in English language teaching and learning in Thailand and also set the CEFR proficiency levels for students to achieve. For English Program students, the goals are: A2 for Prathom 6 (Grade 6), B1 for Matthayom 3 (Grade 9), and B2 for Matthayom 6 (Grade 12).

Consequently, the question whether th+e students in the Program have acquired the desired proficiency level was raised and this led to examining the test results in the Program from the CEPT and OOPT as they are the indicators of the students' proficiency levels according to the CEFR.

Statement and Significance of the Problems

The test results were expected to be the information that the Program could use for making future decisions. However, when the test results were reviewed, it was found that many students received different CEFR levels in each test, resulting in limitations in utilization of the test results. Thus, it was necessary to look at the quality of the information that the tests provided and to hear the opinions from all parties involved in this testing.

Objectives

1. To investigate whether the tests used can provide an accurate indication of the CEFR proficiency levels.
2. To examine the relationship between the results from the Cambridge English Placement Test and the Oxford Online Placement Test.
3. To find out which test result have a higher correlation with student achievement in English courses in the Program.
4. To explore student, teacher, and administrator satisfaction towards this testing, the results, and their intentions of utilizing the information obtained from this testing.

Research Questions

1. To what degree were the CEFR equated levels the students received from both tests consistent?
2. What was the relationship between the results from the Cambridge English Placement Test and the Oxford Online Placement Test?
3. Which test result had a higher correlation with the students' achievement on English courses in the Program?
4. How satisfied were the students, teachers, and administrators towards this testing and the results, and what are their intentions on utilizing the information obtained from this testing?

Literature Review

Proficiency tests

There are many types of tests; however, proficiency tests are the tests used to make program-level decisions, for instance, to set up an entrance or exit standard of a program or to make a comparison with other programs. They are defined as the tests that "assess the general knowledge or skills commonly required or prerequisite to entry into (or exemption from) a group of similar institutions" (Brown, 2005, p. 8).

Another type of test that is very similar to proficiency tests is the placement tests. It is used for placement when there is a need to group students of the same abilities together in order to have a class that teachers can deal with similar ability students. Placement tests can be designed to place students in many levels such as a specific language program, a language course, or a level of a language course (Brown, 2005). Though proficiency tests and placement tests look very similar, Brown (2005) pointed out that proficiency tests tend to be very general in terms of their content and assess a wide range of abilities, from beginner to native level, while placement tests are designed in a more specific way. For example, the content might be related to the curriculum of that particular language program.

Test reliability

For language tests, the term 'reliability' means the accuracy which is reflected in providing the same results when the measurements are repeated in different situations (Henning, 1987). Bachman (1990) gave an explanation regarding test reliability that test performance is affected by many factors other than the true abilities. Those factors are called 'measurement error" and when the measurement error is minimized, the test's reliability is maximized.

There are many factors that affect language test scores. Bachman (1990) grouped these factors into three categories: test method facets, personal attributes, and random factors. Similarly, Henning (1987) also discussed threats to test reliability. The first is fluctuation in the learner which refers to changes in their ability and psychological or physiological state. The second is fluctuation in scoring caused by intra-rater variance, which is the variance of the agreement among repeated ratings done by a single rater, and inter-rater variance, which is the variance of the agreement among raters. Another threat is fluctuation in test administration which includes test instructions and the administrative environment of the test, as well as the characteristics of the test and responses.

Test validity

According to Messick (1990), "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 5). In short, the term 'validity' refers to "the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure" (Henning, 1987, p. 89).

There are many types of validity. This research addressed content validity, concurrent validity and construct validity. Content validity "is concerned with whether or not the content of the test is sufficiently representative and comprehensive for the test to be a valid measure of what it is supposed to measure" (Henning, 1987, p. 94), while concurrent validity is a kind of empirical validity as the evidence is in numerical form obtained from the coefficient between some external criterion measures (Henning, 1987). This can be done either by

examining the differences in test performance among groups of individuals at different abilities or examining the correlations among various measures of a given ability (Bachman, 1990). Finally, construct validity, Brown (1988) explained that 'construct' is a human trait that can be observed only through tests as it occurs in the brain; therefore, it cannot be seen or touched. Therefore, establishing the construct validity of a test means the demonstration that the test is measuring that particular construct (Brown, 1988). Messick (1990) stated that "construct validity is based on an integration of any evidence that bears on interpretation or meaning of the test scores – including content and criterion-related evidence" (p. 14)

Henning (1987) stated that there are several factors that can reduce the test validity. They are the invalid application of tests, inappropriate selection of content, the way the examinees respond to the test, and the use of invalid construct.

The Common European Framework of Reference (CEFR)

The Common European Framework of References "provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe, 2001, p. 1). The framework describes what language learners have to learn to do and what knowledge or skills they need to have to be able to communicate effectively. It also allows proficiency levels to be measured during the learning process and it provides definitions of each proficiency level (Council of Europe, 2001).

The CEFR categorizes language users into six levels in which descriptors of different language skills and competences are given. There are A1 and A2 for Basic Users, B1 and B2 for Independent Users, and C1 and C2 for Proficient Users (Council of Europe, 2001).

Computer adaptive test

According to Brown (1997) a computer adaptive test has other special characteristics other than general computer-assisted tests as the test items are individualized for each test taker. The test ends when the student's ability level is indicated, and it takes less time and fewer test items to finish the test. Dunkel (1999), also stated that second language computer adaptive testing is "a technologically advanced method of assessment in which the computer selects and presents test items to examinees according to the estimated level of the examinees' language ability" (p. 78).

When taking a computer adaptive test, a test taker starts the first test item which is of a moderate level of difficulty using a computer (Dunkel, 1999). If he or she answers that item correctly, the system will send the next question which is more difficult. On the other hand, if the test taker gives the wrong answer, the system will offer easier items next. The test process continues this way until the system is able to locate the test taker's ability (Dunkel, 1999).

The Cambridge English Placement Test and the Oxford Online Placement Test

The Cambridge English Placement Test (CEPT) is an online adaptive test which can be used to find out the level of English ability of the student, classes or courses that are suitable for the students, and which Cambridge English exam students should aim for. (Cambridge English Language Assessment, n.d.). The test focuses on the skills of Reading, Use of English, and Listening. The test features a variety of accents from a range of English-speaking countries. It takes about 30 minutes to complete the test and the questions formats are both multiple-choice and type-in items. When finished, the results are presented in the form of total score out of 100 and the CEFR level

The Oxford Online Placement Test (OOPT), according to the Oxford University Press (n.d.), in Oxfordenglishtesting.com, the OOPT is also an online adaptive test which is used to place students in the appropriate class and measure student's general language ability according to the CEFR. It consists of two parts: Use of English, in which test takers work on approximately 30 items, and Listening, approximately 15 items. The time allowed is variable between 50-90 minutes up to the test user and the total score is 120.

The English Program at Chonkanyanukoon School

Chonkanyanukoon School is a secondary school under the Secondary Educational Service Area Office 18, the Office of Basic Education Commission (OBEC), Ministry of Education, Thailand. The school started the English Program in 2003. Currently, the English Program has two classes for each level from Matthayom 1 to 6. There are no more than 30 students in a class; therefore, there are about 360 students in the Program every year. There are 17 teachers at the Program. They are Filipino, American, British, and Thai. The curriculum at the Program corresponds to The Basic Education Core Curriculum B.E. (2008). The students study English, math, science, and computer with foreign teachers who use English as the instruction medium.

Related research

The OOPT was found to be less favored than other placement tests in the paper "Placement Testing in an EFL Context" (Brooke, Aden, Al-Kuwari, Christopher, Ibrahim, Johnson & Souyah, 2012). This study evaluated the scoring rubric created to use in the selection of English placement tests for English for Academic Purposes at the University of Calgary-Qatar. The tests that were evaluated using the rubric as criteria were the Accuplacer, Compass, University of Michigan English Placement Test, Password, Oxford Online Placement Test (computer-based), and Oxford Placement Test (paper-based). Scores on the Accuplacer were the highest, while those of the Oxford Online Placement Test (OOPT) were the fifth and of the Oxford Placement Test the last among six. Although the findings of this research did not satisfy the purpose of the researcher in validating the rubric, this study provided a detailed description of the test reviews from the perspective of an EFL context. There were several factors which caused the OOPT to score poorly from using the rubric such as the lack of control in the test room, the content that covered the situations that were unlikely for the students to encounter, and heavy emphasis on idioms and British expressions.

Another research study questioning quick placement tests by Berthold (n.d) addressed several issues negatively. The first was the fact that the main component of quick placement tests, whether paper-based or computer-based, is multiple-choice questions. Therefore, there is guessing involved. This study also compared the scores of the paper-based Oxford Placement Test and the Oxford Online Placement Test (OOPT) in the 'Use of English' part. The researcher found that these two tests did not provide consistent CEFR levels to the students.

Research Methodology

This study employed sequential mixed methods where the quantitative phase was conducted first, followed by the qualitative method.

Quantitative method

The results from the CEPT and OOPT, and the achievement scores on English courses were collected from 336 students at the Program (The achievement score was the average of the scores from all English courses that the students took in the second semester of 2013). The data were analyzed by using descriptive statistics and the Pearson Product Moment Correlation coefficient.

Qualitative method

Ten students, three teachers, and one administrator at the Program were interviewed individually using semi-structured interviews. The data were analyzed by the method of coding.

Results

The consistency of the CEFR levels equated levels the students received from the CEPT and OOPT

The study revealed that the CEPT and OOPT did not provide consistent CEFR levels to the students. There were more than approximately 50% of the students who received higher CEFR levels from the CEPT in every Matthayom level. The percentage was particularly high in Matthayom 6, in which there was 61.54%. About 40% of the students in every Matthayom level received the same CEFR levels from the CEPT and OOPT, and there was only one student from every Matthayom who scored in the higher level in the OOPT.

The statistics also showed that the number of students in each CEFR level from the CEPT and OOPT was different. From Mattayom 1 to 6, in the CEPT, A2 was the level that had the greatest number of students, which was 155, followed by B2 at 141 students, with a small number of students in other CEFR levels. However, in the OOPT, the number of students who were in A2 was 226. There were 65 students in A1, and 38 in B1, with a few students in other levels.

Moreover, the average CEFR levels from each test were also different. In the CEPT, the average of every Matthayom was A2, except for Matthayom 4 and 6, which was B1. Nevertheless, in the OOPT, the average CEFR levels of every Matthayom were A2.

The correlation between the results from the CEPT and OOPT

The Pearson Product Moment Correlation coefficient was used to find the correlation between the scores from the CEPT and OOPT. The study found that = .69 at $p < .01$. Therefore, it can be concluded that there was a positive relationship between the scores of the CEPT and OOPT.

The correlation between the students' scores from the tests and their achievement on English courses at the Program

The scores from the OOPT had a higher correlation with the students' achievement on English courses in every Matthayom level. However, the correlations of Matthayom 6 were not significant.

Satisfaction towards this testing and its results

According to the interviews, the informants reported that they were satisfied with the budget and time spent in the test administration, and the test takers felt that the online tests were convenient and easy to complete. Furthermore, the students were satisfied to know about their English proficiency levels.

Dissatisfaction towards this testing and its results

The students were unwilling to take the tests due to not being informed about the purpose and there were also false perceptions of the tests because the students and teachers did not have adequate information about the tests. Some informants expressed that online tests were not a sufficient tool to indicate English proficiency. Moreover, the students commented that the test content was quite different from what they have learned at school, especially the vocabulary and idioms. The test takers also reported that there were distractions in the test room.

After the test results were known, the teachers expressed their concerns towards the low proficiency and there were confusion and doubt in the inconsistency of CEFR levels from both tests among the stakeholders.

The intention on utilizing the test results

The test results were expected to be used in making both program-level and classroom-level decisions. For the program-level decisions, the test results could be used to adjust the English curriculum and course syllabi at the Program. In the classrooms, teachers could use the test results as a benchmark for their courses objectives and as a diagnostic tool.

The test results were also expected to help the students in self-improvement as it could indicate the areas that need to be improved and could be used to predict the results of other tests which are linked to the CEFR.

Limitations of test results utilization

Although the test results were expected to be utilized, there were limitations in test results utilization which were the lack of information and understanding about the testing and CEFR of the stakeholders, as well as the decreased confidence in the quality of the tests.

Discussion and Conclusions

Discussion

This part discusses the results from the quantitative and qualitative phase, followed by the discussions on other issued discovered.

Discussion on the quantitative findings

There were some factors which could have contributed to the inconsistency in the CEFR levels between the CEPT and OOPT besides the content of the tests, such as the differences in the full score and the range of each level, the time allowed and the test parts. Although the scores from the CEPT and OOPT were not consistent, the tendency of the difference in the CEFR levels between the two tests was similar in every Matthayom level. There was also a strong correlation between the results of the CEPT and OOPT. These indicated that the results from one test could predict the results in the other. The lower results from the OOPT were in accordance with the interviews in which the informants said that the OOPT was more difficult than CEPT.

The correlation between the test scores and the achievement on English courses of Matthayom 6 students was not significant. This might have been because the coming admission exam distracted them from doing the tests and the grades at school were given more leniently than to the students in other levels.

Discussion on the qualitative findings

According to the interviews, there were many factors which affected the test results negatively. There was a reduction of the test reliability due to the distractions in the test

rooms and personal attributes of the test takers. The lack of the information regarding this testing also resulted in an unwillingness to take the tests and inadequate involvement from other stakeholders. Moreover, the interviews revealed that the stakeholders perceived the CEPT and OOPT to be low-stake tests.

Test validity

This study did not intend to indicate the validity of the CEPT and OOPT themselves; however, the adoption of the CEPT and OOPT for the purpose of measuring the English proficiency of the students at the Program seemed to lack validity. Firstly, it has an issue regarding construct validity as the tests used contained only receptive skills, but the CEFR contained receptive, productive, and interactive skills. Furthermore, the CEPT and OOPT were also created for placement purposes, not necessarily for indicating one's proficiency. The quantitative findings also showed that there were differences in the scores from the CEPT and OOPT for the same individuals; therefore, these tests lacked concurrent validity.

Conclusions

From the findings, the researcher concluded that, first; the English proficiency level of the students at the Program was lower than expected as the average score was in A2. Second, the CEPT and OOPT might not be valid for proficiency testing at the Program and should be supplemented with other forms of evaluation. Third, the lack of information and understanding of the stakeholders about the testing negatively affected the utilization of the test results. Finally, the CEPT and OOPT were perceived to be low-stake tests which created no washback.

Recommendations and suggestions for further studies

The test users should understand their purpose of testing and having a thorough understanding about the tests before adopting them and online tests should not be the only tool to indicate the students' proficiency, especially when the test results were expected to be used in making decisions. The distractions in the test room should also be minimized.

The Program should investigate the coherence between the expected outcomes which were based on the CEFR and the current English curriculum which was based on the national core curriculum, because the content in the CEFR and the core curriculum were different. Also, the Program should study other tests or measurements for future testing as well as how to minimize the factors that affected the test reliability and validity.

* Ratima Sinlapachai, *E-mail: diorama1986@hotmail.com*

* Janpanit Surasin, *Faculty of Education, Burapha University, Thailand, E-mail: janpanit.s@gmail.com*

* Raweewan Angkanurakbun, *Faculty of Education, Burapha University, Thailand.*

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Berthold, M. (n.d.). Reliability of Quick Placement Test: *How much faith can we place on quick paper or internet based placement tests?*. Retrieved from http://humanities. curtin. edu.au/schools/EDU/education/pdf/alaa/M-Berthold.pdf

Brooke, K., Aden, M., Al-Kuwari, N., Christopher, V., Ibrahim, M., Johnson, B., & Souya, O. (2012). Placement testing in an EFL context. *TESOL Arabia Perspective, 19*(2), 13-20.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59.

Brown, J. D. (2005). *Testing in Language Programs: A comprehensive guide to English language assessment*. Virginia: Prentice Hall Regents.

Cambridge English Language Assessment. (n.d.). *Placing students in the right exam.* Retrieved from http://www.cambridgeenglish.org/find-a-centre/exam-centres/ support-for-centres/placing-students-in-the-right-exam/

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Retrieved from http://www.coe.int/t/dg4/linguistic/ Source/ Framework_EN.pdf

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

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Los Angeles: Newbury House.

Messick, S. (1990). *Validity of test interpretation and use*. Retrieved from http://files.eric. ed.gov/ fulltext/ED395031.pdf

Ministry of Education. (2014, 14 January). *Reform policy on teaching and learning English.* Ministry of Education announcement.

Office of Basic Education Commission. (2013, 12 June). *English proficiency test according to CEFR.* Instruction Number moe04229/183.

Oxford University Press. (n.d.). *Oxford online placement test.* Retrieved from https://www. oxfordenglishtesting. com/defaultmr.aspx?id=3048

Appendix A

The number of students who received higher, equal, and lower CEFR levels from the CEPT and OOPT

Table 1 The number of students who received higher, equal, and lower CEFR levels from the CEPT and OOPT

CEFR levels	Number of students						
	M1 (N=59)	M2 (N=57)	M3 (N=56)	M4 (N=55)	M5 (N=57)	M6 (N=52)	Total (N=336)
CEPT > OOPT	31 (52.54%)	32 (56.14%)	30 (53.57%)	31 (56.36%)	31 (54.39%)	32 (61.54%)	187 (55.82%)
CEPT = OOPT	27 (45.76%)	24 (42.11%)	25 (44.64%)	23 (41.82%)	25 (43.86%)	19 (36.54%)	143 (42.69%)
CEPT < OOPT	1 (1.69%)	1 (1.75%)	1 (1.79%)	1 (1.82%)	1 (1.75%)	1 (1.92%)	6 (1.79%)

Appendix B

CEFR levels of all students from the CEPT and OOPT

Table 2 CEFR levels of all students from the CEPT and OOPT

CEFR Levels	CEPT	OOPT
C2	0	0
C1	4	0
B2	19	6
B1	141	38
A2	**155**	**226**
A1	17	65
Pre A1	0	1

Appendix C

The average score and CEFR levels of M.1 to 6 in the CEPT and OOPT

Table 3 The average score and CEFR levels of M.1 to 3 in the CEPT and OOPT

	M1		**M2**		**M3**		**M1-3**	
	CEPT	OOPT	CEPT	OOPT	CEPT	OOPT	CEPT	OOPT
μ	38.75	27.02	37.3	25.46	37.71	28.39	**37.93**	**26.95**
σ	11.29	11.10	13.07	10.27	12.87	10.45	**12.36**	**10.63**
Average CEFR	A2+	A2	A2+	A2	A2+	A2	**A2+**	**A2**
	M4		**M5**		**M6**		**M4-6**	
	CEPT	OOPT	CEPT	OOPT	CEPT	OOPT	CEPT	OOPT
μ	41.47	31.67	39.88	29.61	42.17	30.06	**41.14**	**30.45**
σ	11.87	13.11	14.65	14.04	13.96	11.55	**13.5**	**12.94**
Average CEFR	B1-	A2	A2+	A2	B1	A2	**B1-**	**A2**