Linking an English Proficiency Test to the CEFR: Setting Valid Cut Scores

Pathamarat Nakanitanon
Department of Western Languages, Chiang Mai Rajabhat University
202 Chang Puak Road, Tambon Chang Puak, Maung, Chiangmai, Thailand 50300
Email: pathamarat@cmru.ac.th

Received: 20-08-31 Revised: 20-11-03 Accepted: 20-12-29

Abstract

The Common European Framework of Reference for Languages (CEFR) is widely used as a reference in EFL English education, and therefore linking English proficiency tests to the CEFR is imperative if test scores are interpreted according to the CEFR levels. This study aimed to align a FRELE-TH-based test, an English proficiency test developed by Chiang Mai Rajabhat University in Thailand, with the CEFR using the Yes/No Angoff method to derive cut scores. The participants were nine university English lecturers purposively selected as judges. These judges made three rounds of consideration regarding the possibility that a borderline test-taker of each CEFR level would correctly answer the test questions. Their judgments, 1 for 'Yes' and 0 for 'No', were then calculated for the cut scores for A1, A2, B1, B2 and C1 which were 22, 36, 57, 80 and 105, respectively. The test scores can now be interpreted in relation to the CEFR, and this meaningful interpretation is useful for further enhancement of students' English abilities. For further study, triangulation of data using a different but appropriate standard setting method should be undertaken to increase validity of the derived cut scores.

Keywords: CEFR, linking, valid cut scores, Yes/No Angoff method

Introduction

Since its establishment in 2001 (Council of Europe, 2001), the Common European Framework of Reference for Languages (CEFR) has been widely used, and a number of studies on mapping high-stake standardized tests onto the CEFR have been conducted in order to link the test scores to the 6 CEFR levels, namely, A1, A2, B1, B2, C1 and C2. Examples of these studies include TOEFL iBT (Tannenbaum & Wylie, 2008), TOEFL Junior Comprehensive Test (Tannenbaum & Baron, 2015), and TOEIC (Tannenbaum & Wylie, 2019). One of the most frequently used methods remains the Yes/No Angoff.

In Thailand, a large number of English proficiency tests have been developed, however, only a small number of studies on mapping test scores to the CEFR have been undertaken, one of which pertains to the alignment of Srinakharinwirot University Standardized English Test (SWU-SET) with the CEFR. In this study, the SWU-SET was developed and mapped to the CEFR using the Angoff method, and the cut scores of the SWU-SET for A2, B1, B2 and above resulted in 22, 50, and 78 points, respectively (Arthiworakun, Vathanalaoha, Thongprayoon, Rajprasit, & Yaemtui, 2018). Another research undertaken by Wudthayagorn (2018) aimed to map the CU-TEP to the CEFR using an extended form of Angoff called the Yes/No Angoff method. In this research thirteen experts decided whether or not a borderline test-taker of each CEFR level would correctly do the CU-TEP items. The cut scores of the CU-TEP for A2, B1, B2 and C1 were 14, 35, 70 and 99 points out of 120 point-scale, respectively (Wudthayagorn, 2018).

As for Chiang Mai Rajabhat University (CMRU), testing is conducted not only to reflect the students' English language ability for further development but also to certify their English proficiency prior to graduation. The university has therefore developed a test called Chiang Mai Rajabhat University Test of English Proficiency (CMRU-TEP) based on the Framework of Reference for English Language Education in Thailand (FRELE-TH), the framework of reference adapted from the CEFR by the Chulalongkorn University Language Institute (CULI) and the Language Institute of Thammasat University (LITU) and funded by the Thailand Professional Qualifications Institute (TPQI) (Hiranburana, Subphadoongchone, Tangkiengsirisin, Phoochaeoensil, Gainey, Thogsongsri, Sumonsriworakun, Somphong, Sappapan, & Taylor, 2018). Even though the CMRU-TEP was developed based on the FRELE-TH, the CEFR-based framework, mapping the test scores onto the CEFR levels had not been undertaken, and the test scores could not be interpreted with respect to the CEFR. In order to solve this problem, a standard setting study to align the CMRU-TEP scores with the CEFR levels was required. The cut score ranges derived in this study contribute to a meaningful interpretation of the CMRU-TEP scores in relation to the CEFR levels, and the CMRU-TEP can be used as a mirror which reflects students' English proficiency with respect to the CEFR. The test results then can be used to certify the English proficiency of the CMRU undergraduate and graduate students. In addition, these test results would be beneficial for policy makers to make an administrative decision on what and how to do to elevate the students' English language proficiency to national required levels.

Literature Review

Two main related literature reviewed in this section include standard setting and the Yes/No Angoff method.

Standard Setting

In order to use any tests as tools to identify test-takers' performance, it is necessary to set standards to give meaning and relevance to the test scores. According to Hambleton, Jaeger, and Plake (2000), in education, the interpretation of criterion-referenced test score requires specific sets of performance standards. Standard setting is the process to establish cut scores in order to classify the levels of test-takers' performance (Cizek, 2012). It involves human judgments in the process as Livingston and Zieky (1982, p. 12) states that "[a]ny standard—absolute or relative—is based on some type of judgment." While over 60 methods of standard setting have been identified (Kaftandjieva, 2010), the distinctive basic methods based on judgments include Nedelsky, Angoff, and Ebel (MacCan & Gordon, 2004). Of all the three basic methods, Angoff method is more frequently employed (Sireci, Robin, & Potelis, 1999) since it is easier to implement when compared to the others. According to Livingston and Zieky (1982), this method requires that judges make consideration on the probability (0.00-1.00) that a test-taker is able to correctly do each test question, and the mean score is then computed from the probability values. This standard setting method is appropriate for use with multiple choice tests and other types of tests, and it focuses on probability of producing correct answers. Even though the traditional Angoff method is easy to practice, estimating probability of correct answers is difficult to some judges (Impara & Plake, 1997). Therefore, some extensions to this method have been introduced, one of which commonly used in studies, which pertain to aligning a test with the CEFR is the Yes/No Angoff (Tannenbaum & Baron, 2015).

The Yes/No Angoff Method

This method is widely used in standard setting to link the high-stake standardized tests with the CEFR, which include such tests as TOEFL iBT (Tannenbaum & Wiley, 2008), TOEFL ITP (Tannenbaum & Baron, 2011), TOEFL Junior Comprehensive Test (Tannenbaum & Baron, 2015), and TOEIC (Tannenbaum & Wylie, 2019). According to Impara and Plake (1997), similar to the traditional Angoff method, the basic steps of the Yes/No Angoff method include selecting qualified judges, judges' making judgment on the borderline test-takers, averaging the judgment scores, followed by analysing the data for mean cut score and finally discussing the mean cut score and agreeing upon the cut score obtained. However, the Yes/No Angoff is different from the traditional Angoff method in that, instead of correctly estimating probabilities, judges simply make yes/no estimates, giving 1 score for "Yes" and 0 for "No". Thus, instead of making correct estimate proportion, judges would conceptualize a real test-taker they know. The cut score is obtained from averaging judgment scores (Impara & Plake, 1997).

Methodology

Instruments

Judges in the cut score deriving step

Careful selection of judges in standard setting based on judgments is essential. According to Fulcher (2010), more qualified and experienced judges assure a better process of standard setting and validity of judgments. Regarding the number of judges, even though a higher number is preferable (Livingston & Zieky, 1982), six to nine is a common number (MacCann & Gordon, 2004). In some cases, as few as five is used; however, the research results are to be taken as a recommendation only (Livingston & Zieky, 1982). For standard setting of some high-stakes tests, it is reasonable to involve a larger number and variety of stakeholders. For instance, in connecting the TOEFL iBT to the CEFR levels, 23 judges from various positions and geographical areas were involved (Tannenbaum & Wylie, 2008), and 22 judges were recruited for linking the TOEFL ITP and the CEFR. In this particular study, nine judges, who are university English lecturers, participated in the process. The number of judges was appropriate for the context of this study, and it was in accordance with the recommended number of participants—six to nine (MacCann & Gordon, 2004). The judges' average years of teaching experience was 14.3. They were purposively selected using the following criteria:

(1) having at least 10 years of experience in teaching English at the university where the test-takers studied so that they were familiar with the teaching and learning context and had accuracy in judging students' language abilities,

(2) having been involved in the development of the English proficiency test at that university so that they were familiar with the test items and understood the objective of each test item very well, and

(3) understanding the descriptors of the CEFR levels so that they did not have difficulty making judgments.

The CEFR global scale

The CEFR global scale consists of three standard levels of users: basic users (A1 and A2), independent users (B1 and B2), and proficient users (C1 and C2) (Council of Europe, 2001). As stated earlier, the CEFR levels used in this study included A1, A2, B1, B2 and C1, and the descriptor of each level indicates what the test-takers can achieve in terms of the language use. Level C2 was excluded from this study as it was not appropriate for use to assess the English proficiency of the target test-takers who were undergraduate and graduate students in the context of this study.

A FRELE-Based test*: *The CMRU-TEP

The English proficiency test in this study was the Chiang Mai Rajabhat University Test of English Proficiency (CMRU-TEP), a test used for assessing the CMRU students' English proficiency prior to their graduation. Based on the FRELE-TH global scale, comprising of 10 levels of competency in English, namely, A1, A1+, A2, A2+, B1, B1+, B2, B2+ C1 and C2 (Hiranburana et al., 2018; Hiranburana, 2020), the CMRU-TEP was developed to be a multiple-choice test consisting of 120 items divided into three sections—listening, reading and writing according to the descriptors of the FRELE-TH global scale for levels A1, A1+, A2, A2+, B1, B1+, B2, B2+ and C1 only. Time allotted for completing the test tasks was 2½ hours. The test form used in this study had high content validity (0.95) and reliability (0.90).

Judgment form and judgment compilation form

Three sets of expert judgment forms were employed for data collection for the three rounds of judgments pertaining to the five respective levels according to the CEFR. The judgment compilation forms were then used by the researcher to record scores produced by the expert judgments.

Interview points

The two main points asked the participants to include the perceptions on the score mapping process through the Yes/No Angoff method, the CEFR descriptors and problems and solution regarding the test score mapping process.

Data Collection

Mapping the test scores to the CEFR

Before judges made their assessments, an orientation meeting was held to introduce the process of the Yes/No Angoff method and to provide judging tools including judgement forms and the CEFR global scale descriptors to the judges. The researcher allowed time for judges to review the material and to ask questions about the steps in the process, instead of training them in order to prevent any manipulation and to assure accurate and valid judgment. According to Fulcher (2010, p. 244), training can prevent the experts from seeing other possibilities and hence eliminates “the richness of human judgment” and causes reduction of validity. He suggests that instead of training the judges to make judgments, the researcher should develop descriptors with the help of which untrained judges can independently link the descriptors and the performances. Therefore, in this study, the judges considered the ability of a borderline test-taker of each CEFR level based on their own untrained judgments.

Following the Yes/No Angoff process, for three rounds, each judge decided whether or not a borderline test-taker of a CEFR level would answer each test question correctly. Each judge gave one (1) score for “Yes” and zero (0) for “No” for each of the 120 test questions. Interestingly, they agreed to judge basic level (A1-A2) first, then the proficient level (C1), and subsequently the independent (B1-B2) at a later time. They viewed that it was easier to judge the independent levels after finishing the other levels. After completing each round of judgment and calculating for the mean scores, the judges discussed and agreed upon the cut scores. The mean scores of the third round were taken as the final cut scores.

Exploring judges’ opinions on the scores mapping process

The researcher interviewed the judges regarding the following topics:

1) understanding in respect to the purpose of the score mapping through the Yes/No Angoff method;
2) understanding the steps in the score mapping process;
3) understanding the global scale of the CEFR
4) appropriateness in the use of the score mapping method;

5) issues associated with implementing this score mapping and solution;
6) difficulties in making judgments for each level of the CEFR

Data Analysis

The scores derived from the expert judgment were analysed using descriptive statistics including minimum (min), maximum (max), mean ($\bar{x}$), and standard deviation (SD). The standard error of judgement (SEJ) was analysed using Central Limit Theorem (CLT), which is appropriate to use when judgments are made independently (MacCann & Stanley, 2004). The SEJ is important since it indicates the extent of uncertainty in the expert judgments. The formula for this calculation is the standard deviation divided by the square root of number of experts (MacCann & Stanley, 2004; Cizek & Bunch, 2007 as cited in Tannenbuam & Baron, 2015).

Results

Expert Judgments and Cut Score Ranges

This section presents the results of data analysis in four parts: expert judgments for Round 1, expert judgments for Round 2, expert judgments for Round 3, and cut score ranges mapped onto the CEFR levels. The statistical data are presented in Tables 1-5.

Table 1

Expert Judgment for Round 1

CEFR Levels	**Min**	**Max**	$\bar{x}$	**S.D.**	**SEJ**
A1	10	45	22.11	10.61	3.54
A2	23	66	38.33	14.28	7.76
B1	41	92	70.33	17.59	5.86
B2	74	115	95.11	16.86	5.62
C1	92	120	110.22	9.68	3.23

Table 1 shows the results of data analysis of expert judgment for Round 1. The mean scores for A1, A2, B1, B2 and C1 were 22.11, 38.33, 70.33, 95.11 and 110.22, which were rounded to 22, 38, 70, 95 and 110, respectively. B1 had the largest standard deviation (17.59) and B2 the second largest, whereas C1 had the smallest standard deviation (9.68). This means that the judges' opinions on the English ability of the borderline test-takers of C1 were more similar to one another than those of the other CEFR levels. The standard error of judgment of C1 was also the lowest (3.23), indicating that there was a lower rate of uncertainty in the judges' consideration when compared to the other levels.

Table 2

Expert Judgment for Round 2

CEFR Levels	Min	Max	$\bar{x}$	S.D.	SEJ
A1	10	56	22.44	15.92	5.31
A2	19	70	35.78	17.75	5.92
B1	32	78	56.67	18.08	6.03
B2	60	94	79.78	13.38	4.46
C1	70	117	103.22	15.32	5.11

Table 2 presents the results of data analysis of Round 2 judgment. In this round, the mean scores for A1, A2, B1, B2 and C1 are 22.44, 35.78, 56.67, 79.78 and 103.22. The rounded mean scores were 22, 36, 57, 80 and 103, respectively. Overall, the mean scores in Round 2 were smaller than those of Round 1, except for A1, which remained the same (22). The standard deviation and the standard error of judgments of B2 were the smallest (13.38 and 4.46), while those of B1 were the largest in this round (18.08 and 6.03). It can be summarized that the judges had more different opinions for B1 and expressed the highest rate of inconsistency in their judgments when compared to the other CEFR levels.

Table 3

Expert Judgment for Round 3

CEFR Levels	Min	Max	$\bar{x}$	S.D.	SEJ
A1	11	58	22.22	15.21	5.07
A2	24	67	36.22	15.41	5.14
B1	31	84	57.44	22.17	7.39
B2	44	92	79.89	17.16	5.72
C1	78	119	105.35	13.67	4.56

Table 3 shows the results of data analysis of Round 3 judgment. In this round, the mean cut scores for A1, A2, B1, B2 and C1 were 22.22, 36.22, 57.44, 79.89 and 105.33, and they were rounded to 22, 36, 57, 80 and 105, respectively. The cut scores obtained in this round were not much different from those in Round 2. The cut score for A1 remained the same in all three rounds of judgments, and those for B2 were the same in Round 2 and Round 3 (80). This means that there was very much agreement and consistency in judges' considerations. However, the minimum score for B2 in this round was much lower than that of Round 2, and the standard error of judgment was larger in Round 3. Similar to the cut scores of B1 in Rounds 1 and 2, the cut score of B1 in Round 3 had the largest standard deviation and the standard error of judgment (22.17 and 7.39), and the smallest for C1 (13.67 and 4.56), indicating that the judges agreed more on C1 and were the least inconsistent in their judgments. The derived mean cut scores from this round were rounded and put in ranges according to the CEFR as presented in Table 4.

Table 4

Cut Score Ranges of the CMRU-TEP Mapped to the CEFR

CEFR Levels	**A1**	**A2**	**B1**	**B2**	**C1**
CMRU-TEP Scores	22-35	36-56	57-79	80-104	105-120

Table 4 illustrates the CMRU-TEP cut score ranges mapped to the CEFR levels. The CMRU-TEP cut scores were based on the rounded mean scores obtained in Round 3 judgments. The cut score ranges for A1, A2, B1, B2 and C1 were 22-35, 36-56, 57-79, 80-104 and 105-120, respectively. The cut score range for B2 was the widest (24 points), and the range for B1 was the second widest (22 points). All the participating judges agreed with the cut score ranges, which were the output for this standard setting study.

Interview Results

According to the results of the group interview conducted after the standard setting process, it was found that all the judges understood the objective of the standard setting and the process of the Angoff method as they could clearly elaborate on the objective and the steps involved in the process. They also stated that they understood the CEFR global scale; however, when bringing into practice, they found it difficult to judge levels B1 and B2. They stated that making judgments on C1 was the easiest, followed by A1 and A2. B1 was the most difficult level, followed by B2. Therefore, in the second and the third rounds, they decided to judge the levels A1, B1, C1, A2 and B2, respectively.

Discussion

Inconsistent Ranges of Cut Scores

In this standard setting, using the Yes/No Angoff method, it could be observed that the cut score ranges were different from level to level, 13 points for A1, 20 points for A2, 22 points for B1, 24 points for B1, 24 points for B2, and 15 points for C1.This finding was in accordance with the observation in a study conducted by Wudthayagorn (2018) in which the cut score ranges were much different—35 points for B1, 29 points for B2, 22 points for C1, and 21 points for A2. In that study, the cut score ranges for the lowest and the highest scales were the smallest since they were easier to observe and to judge when compared to the middle scale. This consequence was from the purposive design of the CEFR to allow flexibility for any local adaptation of the scales so that they can be applied in multiple contexts and used for all languages (Council of Europe, 2020).

Standard Error of Judgment

Standard setting based on judgments, to some extent, can lead to some errors in making judgments. In order to minimize the rate of errors, the selection of qualified judges should be carried out carefully so that a good implementation process and valid judgments can be expected. According to Fulcher (2010), more qualified and experienced judges can lead to better judging processes. An important reason is that the experienced judges have the ability to assess students' English competencies more accurately. Based on the results of several studies, the accuracy of judgment can be increased through judges' conceptualization of a single real target test-taker known to them (Impala & Plake, 1997). The judges in this study have at least 10 years of experience teaching English to CMRU students, so they have a better conceptualization of a student who is categorized as a borderline test-taker of each CEFR level. Hence, their judgments' inconsistencies were minimized despite the fact that discrepancies between levels had occurred.

Results of data analysis revealed that the error of judgment in all rounds for the C1 level was the smallest, followed by the A1 and A2 levels. This indicates that the judges most consistently shared the common opinions on the C1 borderline test-takers' abilities and A1 as the second most. B1 was the most problematic level and B2 the second most. It can be interpreted that there was some level of inconsistencies in the judges' decision making for the borderline test-takers of the independent levels of the CEFR, B1 and B2. This finding was in accordance with Wudthayagorn's (2018). The unclear-cut boundary design of the CEFR levels makes it difficult to identify the B1-B2 levels, which are in the middle of the scale, and this can cause, to some extent, variances in judgments (Wudthayagorn, 2018). However, this unclear-cut boundary design has a good purpose. According to the Council of Europe (2001), the holistic specification of the CEFR (2001) global scale aims to make it applicable to various contexts and languages. In order to optimally use the CEFR, specific supporting guidelines are required (Foley, 2019), and for this reason, the FRELE-TH was developed to be appropriate for use in the Thai context.

Limitations

A limitation of this study was the characteristics of the CMRU-TEP. The CMRU-TEP assesses only the listening, reading, and writing skills; hence, the interpretation of the test scores in relation to the CEFR levels is only approximate. In addition, due to the time constraint, triangulation of the derived cut scores has not been conducted.

Recommendations

Using the Cut Scores to Identify the CEFR Levels

The cut scores obtained from a standard setting study are essential, as they make the test results more meaningful in that they provide an answer to the question into which CEFR level a test-taker should be placed and how the test results are to be interpreted. Once the test scores are linked with the CEFR, they can be interpreted in relation to the CEFR. The cut scores obtained from this study, for example, can now be interpreted in relation to the CEFR levels A1, A2, B1, B2 and C1, respectively. Now the CMRU-TEP can be utilized to assess the English abilities of CMRU undergraduate and graduate students before they graduate, and the test results will be useful both for certifying the students' English abilities according to the CEFR and for further development of students' English proficiencies.

Further Study

In order to ensure validity of the derived cut scores, a further study to triangulate the data could be undertaken by triangulating the data, which could be achieved by using a different but appropriate method of standard setting with the same group of judges to derive a new set of cut scores. Should the variance between the two sets of cut scores—one obtained from using the Yes/No Angoff and one from using the new method—prove to be insignificant, it will then confirm the validity of cut scores derived from the present study, which utilizes the Yes/No Angoff method.

In addition, a study to examine effectiveness of the use of cut scores in relation to the CEFR should be undertaken in order to ensure that the cut scores have capacity to correctly identify test-takers' English abilities regarding the CEFR levels, and false positive or negative errors are minimized.

Acknowledgements

This research was funded by Chiang Mai Rajabhat University and supported relevant data by the CMRU Language Center.

References

Athiworakun, C., Vathanalaoha, K., Thongprayoon, T., Rajprasit, K., & Yaemtui, W. (2018). SWU-SET as a CEFR standardized English test. *Journal of Language Teaching and Research, 9*(2), 261-267. doi: http://dx.doi.org/10.17507/jltr0902.06

Cizek, G.J. (2012). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G.J. Cizek, (Ed.). *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3-14). NY: Routledge.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment—Structured overview of all CEFR scales.* Retrieved from https://rm.coe.int/168045b15e

Council of Europe. (2020). *Common European Framework of Reference for Languages: The CEFR Levels*. Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Foley, J.A. (2019). Adapting CEFR for English Language Education in ASEAN, Japan and China. *The New English Teacher, 13*(2), 101-117. Retrieved from http://www.assumptionjournal.au.edu/index.php/newEnglishTeacher/article/view/3879

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Hiranburana, K. (2020). FRELE-TH: Springboard for holistic English educational reform.

LEARN Journal, 13(1), 62-75.

Hiranburana, K. Subphadoongchone, P, Tangkiengsirisin, S., Phoochaeoensil, S., Gainey, J., Thogsongsri, J., Sumonsriworakun, P., Somphong, M., Sappapan, P. & Taylor, P. (2018). Framework of reference for English language education in Thailand—(FRELE-TH) based on the CEFR: Revisited in the English educational reform. *Pasaa Paritat, 33*, 51-91. Retrieved from http://www.culi.chula.ac.th/publicationsonline/files/article2/vP11nmOMYxMon45327.pdf

Hoge, R.D. & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297-313. Retrieved from http://www.jstor.org/stable/1170184

Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*(4), 353-366. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1745-3984.1997.tb00523.x

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL.* Arnhem, The Netherlands: CITO. Retrieved from https://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Livingston, S. A. & Zieky, J. M. (1982). *Passing scores: A manual for setting standards of performanceon educational and occupational tests*. Princeton, NJ: Educational Service. Retrieved from https://www.ets.org.Media/Research/pdf/passing_scores.pdf

MacCann, R. G. & Gordon, S. (2004). Estimating the Standard Error of the Judging in a modified-Angoff Standards Setting Procedure. *Practical Assessment, Research, and Evaluation*, *9*, 1-9. https://doi.org/10.7275/n78q-6g60

Sireci, S.G., Robin, F. & Potelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education*, *12*(3), 301-325, DOI: 10.1207/S15324818AME1203_5

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping TOEFL ITP scores onto the Common European Framework of Reference*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RM-11-133.pdf

Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping scores from the TOEFL Junior® Comprehensive test onto the Common European Framework of Reference (CEFR)*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RM-15-13.pdf

Tannenbaum, R. J. & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology*. Princeton. NJ: ETS. Retrieved from https://www.ets.org/RR-08-34.pdf

Tannenbaum, R. J. & Wylie, E. C. (2019). *Mapping the TOEIC® Tests on the Common European Framework of Reference*. Retrieved from https://www.ets.org/s/toeic/pdf-cefr-flyer.pdf

Wudthayagorn, J. (2018). Mapping the CU-TEP to the Common European Framework of Reference (CEFR). *LEARN Journal*, *11*(2), 163-180. Retrieved from http://www.so04.tci-thaijo.org/index.php/LEARN/article/view/161641