Nov 30, 2022

Aligning the CEFR Levels with the TU-GET CBT Scores

Principal Investigator (PI): Sun-Young Shin[1]
Co-PI: Suchada Sanonguthai[1] and Supong Tangkiengsirisin[2]

[1] Indiana University
[2] Thammasat University

The research ethic committee of Thammasat University approved this study protocol (IRB approval number: 084/2465)

The Thammasat University General English Test (TU-GET) Computer-based test (CBT) is a standardized general English proficiency test required for Thammasat University's undergraduate and graduate students. It can also be taken by people who are interested in assessing their overall English competency. The test battery of the TU-GET CBT consists of four main parts: listening (30 questions), reading (30 questions), speaking (1 task), and writing (1 task). All of the test items in the TU-GET CBT listening and reading sections are based on the multiple-choice format with four options. The TU-GET CBT speaking section consists of one opinion speech task and the TU-GET CBT writing section consists of one essay writing task. Each section has a score range of 0-30. These are added together to a total score of 0-120.

This report includes the results of an empirically-evidenced linking study to align the TU-GET CBT scores to the widely used Common European Framework of Reference (CEFR; Council of Europe, 2001) levels by employing two commonly used standard-setting methods: the Yes/No Angoff method (Impara & Plake, 1997; Hsieh, 2013) and the Bookmark method (Karantonis & Sireci, 2006; Shin & Lidster, 2017). Note that this is not the first time for the Language Institute of Thammasat University (LITU) to align TU-GET CBT scores to other English language proficiency test scores such as TOEFL-iBT and IELTS (See the website link[1] in the endnote for more information). However, given the growing use of TU-GET CBT for matriculated undergraduate and graduate students at Thammasat University (TU) and beyond, it is opportune for the TU-GET CBT test scores to be mapped onto the CEFR levels through rigorous evidence-based procedures to enhance interpretability and meaningfulness of the TU-GET CBT scores for all involved stakeholders.

Aligning test scores with external criteria of language proficiency through a standard-setting approach (Tannebaum & Cho, 2014; Shin & Lidster, 2017) provides useful information about where test takers stand in relation to their overall language learning process because a test score by itself does not directly indicate a proficiency level. Among existing language proficiency guidelines, the CEFR is chosen in this study because it is most commonly used in the

context of foreign language teaching and assessment, representing a progression of language proficiency in six levels (Shin, 2013).

This study was conducted to identify four distinct cut scores for each subsection of the TU-GET CBT which will be used to link each subsection and total TU-GET CBT scores to five CEFR levels (A1, A2, B1, B2, and C1/C2). The distinction between C1 and C2 levels was not made due to the limited number of test items and the paucity of the target population who could belong to the C2 level. The whole project was led and guided by PI, Dr. Sun-Young Shin, but TU-GET CBT test data was prepared and analyzed by a Co-PI (Aj. Suchada Sanonguthai). And a panel of 12 experts who are familiar with TU-GET CBT and the target test taker population were recruited and organized by another Co-PI (Dr. Supong Tangkiengsirisin).

This standard-setting study was conducted entirely online via Zoom due to the COVID-19 pandemic.12 panelists who were familiar with the TU-GET CBT and the target population were recruited. 12 faculty members from LITU and other universities with Ph.D. degrees[2] and with extensive English teaching experiences served as panelists for the standard setting because a minimum of 10 panelists representing broad and diverse backgrounds is suggested in the standard-setting literature (see Loomis, 2012; Tannebaum & Cho, 2014). Each panelist was provided with documents for the abridged CEFR level descriptors for each language skill and the overview of TU-GET CBT.

During the first Zoom meeting, Dr. Shin introduced the CEFR level descriptor to the panelists to ensure that each panelist has a common, agreed-upon understanding of the CEFR level descriptors for four English language skills, and he also presented information about the TU-GET CBT in detail with sample passages and items to familiarize panelists with the TU-GET CBT test content. A practice session was then provided for the subsequent yes/no Angoff standard-setting and Bookmark methods to identify the cut scores for each level using the sample TU-GET CBT items.

TU-GET CBT Reading and Listening Section Cut Scores

Both the Yes/No Angoff and the Bookmark methods were used to generate four cut-scores. 6 panelists used the Yes/No Angoff method and the other 6 panelists employed the Bookmark method and subgroup of 3 panelists for each standard-setting method was formed and the four group leaders were selected.

In the Yes/No Angoff method, 6 panelists conceptualized a borderline test taker who possess a minimum level of four CEFR levels (A2, B1, B2, and C1) and then judged whether the borderline test takers at each level would answer each item correctly (Yes equal to 1) or incorrectly (No equal to 0). This Yes/No Angoff method is appropriate for generating multiple cut scores and is also known to be less cognitively demanding to panelists while yielding comparable cut scores obtained by a traditional Angoff method (Plake & Cizek, 2012).

In the first round of the Yes/No Angoff standard-setting procedure, panelists completed their individual yes/no judgments for all items. Their individual four cut scores were compared with others and received feedback about their judgments. As part of feedback and panel discussion in each subgroup, information about overall item difficulty and item difficulty of the top 25% and bottom 25% of test takers was provided to panelists to guide their cut score setting. After discussion, panelists resumed their ratings for each item in the second round. After the second round of ratings, Dr. Shin and Aj. Suchada examined the levels of agreement on each cut score generated from two subgroups. Afterward, panelists repeated the ratings for each item in the third round. After the third round in which each group leader discussed the results with Dr. Shin and Aj. Suchada, the median of panelists' cut scores was determined as the final cutoff point for each level. Table 1 below shows the cut scores for each four CEFL levels generated by the Yes/No Angoff method.

Table 1. TU-GET CBT Reading and Listening Cut Scores for CEFR levels by the Yes/No Angoff method

CEFR levels	Reading	Listening
A1/A2	5 below	9 below
B1	6 - 14	10 - 17
B2	15 - 25	18 - 26
C1/C2	26 above	27 above

The Bookmark method groups followed a similar procedure, although they received an Ordered Item Booklet (OIB) determined by the Item Response Theory (IRT) technique (Karantonis & Sireci, 2006). This analysis utilized the data which was collected from multiple TU-GET CBT official administrations between November 2020 and February 2022 (N = 479). The OIB contains the items ordered by difficulty from easiest to most difficult and each panelist is asked to place the bookmark between each CEFR level at the last item where students who are assumed to be minimally competent for each level would answer it correctly higher than or equal to 2/3 (approximately 67%). As with the Yes/No Angoff method, panelists received feedback about their bookmark placement for each CEFR level and discussed their bookmark locations and item statistics including item difficulty and item discrimination values in addition to the characteristics of borderline students between each round. The same procedure was repeated over three rounds. Table 2 displays the cut scores for each four CEFR levels generated by the Bookmark method.

Table 2. TU-GET CBT Reading and Listening Cut Scores for CEFR levels by the Bookmark method

CEFR levels	Reading	Listening
A1/A2	4 below	5 below
B1	5 - 13	6 - 13
B2	14 - 23	14 - 24
C1/C2	24 above	26 above

Table 3 below shows the final cut scores based on the median cut scores generated by both the yes/no Angoff and the Bookmark standard-setting methods. Note that there are no A1-level items present in both Reading and Listening sections, which disallowed panelists to judge the probability of marginal A1 students to answer the items more than two-thirds of the time.

Table 3. Final TU-GET CBT Reading and Listening Cut Scores

CEFR levels	Reading	Listening
A1/A2	5 below	7 below
B1	6 - 14	8 - 15
B2	15 - 24	16 - 25
C1/C2	25 above	26 above

TU-GET CBT Speaking and Writing Section Cut Scores

Table 4 displays the cut scores of TU-GET CBT speaking and writing sections for CEFR levels. These cut scores were generated by a performance profile approach (Hambleton et al, 2000) in which a set of borderline profiles of performance scores across possible ranges of scores is presented to panelists, and classification of borderline profiles of performances is made according to their agreed or established standards. The final cut score is determined by the median scores of panelists' ratings of each borderline performance.

Table 4. TU-GET CBT Speaking and Writing Cut Scores for CEFR levels

CEFR levels	Speaking	Writing
A1	1 - 3	1 - 3
A2	4 - 8	4 - 9
B1	9 -15	10 - 15
B2	16 - 23	16 - 22
C1/C2	24 above	23 above

12 panelists rated 30 each of TU-GET CBT speaking and writing examples in the first round after they were calibrated on the CEFR level descriptors and speaking and writing samples that typified each CEFR level. Multiple speaking and writings samples collected from the previous TU-GET administrations that received the scores ranging from 1 to 30 were reviewed and selected by Dr. Shin. Note that these examples of speaking and writing were used anonymously, and panelists did not receive any personal information about test takers. The inter-rater reliability was .976 for the speaking section and .967 for the writing section and intra class correlation coefficient was .768 for speaking and .710 for writing sections in the first-round rating. Dr. Shin had an online Zoom meeting with panelists after the first round and provided feedback and comments on panelists' first ratings and discussed the samples that they disagreed with each other most. In the second round, the inter-rater reliability was .980 for the speaking section and .981 for the writing section and intra class correlation coefficient was .806 for speaking and .811 for writing sections which are considered high reliability. After the second-round rating, Dr. Shin met with four group leaders from every four subgroups and finalized ratings of a couple of samples that panelists still disagreed with each other.
Table 5 below is the final total TU-GET CBT cut scores for CEFR levels which combine cut scores from each four TU-GET CBT section.

Table 5. Total TU-GET CBT Cut Scores for CEFR levels

CEFR levels	TU-GET Scores
A1	1 - 19
A2	20 - 32
B1	33 - 62
B2	63 - 97
C1/C2	98 above

This report contains the CEFR cut scores for each four subsection and total scores of the TU-GET CBT which are recommended by 12 panelists with the guidance of Dr. Shin and Aj. Suchada through rigorous three-round ratings. This collaborative standard-setting study provides empirical evidence on the degree to which the TU-GET CBT test scores relate to the CEFR levels. It can also enhance the interpretability and meaningfulness of the TU-GET CBT test scores and help the TU-GET CBT to achieve wider recognition in both Thailand and overseas.

For future direction, it would be recommended to compare these cut scores with the recent TU-GET CBT scores of test takers who are identified as being borderline students by their teachers who are familiar with them. This further evidence would corroborate the cut scores yielded from the Yes/No Angoff and the Bookmark methods.

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the CEFR: Learning, Teaching, and Assessment.* Strasbourg, France: Council of Europe.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355-366.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47-76). New York, NY: Routledge.

Hsieh, M. (2013). Comparing Yes/No Angoff and Bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10*(3), 331-350.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353-366.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice, 25*, 4–12.

Loomis, S. C. (2012). Selecting and training standard setting participants: State of the art policies and procedures. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 107-134). New York, NY: Routledge.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and Yes/No standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181-199). New York, NY: Routledge.

Shin, S.-Y. (2013). Proficiency scales. In C. A. Chappelle (Ed.), The Encyclopedia of Applied Linguistics (pp.1-7). Oxford, UK: Wiley-Blackwell.

Shin, S.-Y., & Lidster, R. (2017). Evaluating standard setting methods in an ESL placement testing context. *Language Testing*, *34*(3), 357-381.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*(3), 233-249.

Notes

1. http://public.litu.tu.ac.th/view/post/37
2. All 11 panelists have a Ph.D. degree and one non-LITU panelist is currently a Ph.D. candidate.