

Contrasting Groups Analysis of *TOEFL® iBT* Test Cut Scores and the Common European Framework of Reference (CEFR) Proficiency Levels: Kernel Density Estimation of an English Learners' Corpus

Peter Kim^{1*}

Received: 7 January 2021

Accepted: 15 February 2021

Abstract

Placing non-native speakers of English into appropriate classes involves mapping placement test scores onto proficiency levels based on predetermined cut scores. However, studies on how to set boundaries for different levels of proficiency have been lacking in the language testing literature. A top-down approach to standard setting in which a panel of experts set cut scores has dominated the typical standard setting procedure. A less utilized approach is to proceed bottom-up by clustering learners based on test scores. The purpose of this study was to fill this gap by examining Education Testing Services (ETS)'s mapping of *TOEFL® iBT* Test scores to the Common European Framework of Reference (CEFR) levels. The study examined *TOEFL® iBT* score data from ICNALE (International Corpus Network of Asian Learners of English) and conducted optimal Kernel Density Estimation to find peaks in the distribution of test scores. In addition to the number of peaks, the local minima of the resulting distribution were chosen as boundaries of cut scores for delineating different ability groups. This method of separating scores, also known as contrasting groups, finds clusters of test takers based on maximum differences in scores. The results showed that ETS' guide for cut scores linking to CEFR levels was comparable to Kernel Density Estimation with some exceptions, namely two out of three cut scores were found to be similar. Implications are discussed in terms of test-centered versus examinee-centered method of standard setting and the need to consider the demographics of the examinee population in determining cut scores.

Keywords: CEFR; cut scores; proficiency levels; standard-setting; TOEFL

1. Introduction

One of the most commonly accepted guidelines for determining proficiency levels of foreign languages is the Common European Framework of Reference for Languages (CEFR). CEFR is an internationally recognized set of standards used for describing language proficiency. It started in 1990 by the Council of Europe as a wider effort to promote collaboration between language teachers across all European countries by providing clear descriptions needed to

¹ Teachers College, Columbia University * (corresponding author: pk2505@tc.columbia.edu)

evaluate learners' language proficiency for purposes of both teaching and assessment, particularly in an academic setting. CEFR uses an action-orientated approach to describing language learning outcomes and divides language activities into four skills of listening, speaking, reading and writing (Little, 2006; Papageorgiou et al., 2015; Tannenbaum & Wylie, 2008). Qualitative descriptors of these skills – what language learners are able to do – are used to define six common reference levels of A1, A2, B1, B2, C1, C2 (For a description of these levels see Appendix A). Level A is considered basic user (Beginner), level B is considered independent user (Intermediate) and level C is considered proficient user (Advanced) (see Appendix A for details). The three major levels of A, B, and C are further divided into A1, A2, B1, B2, and C1, C2, in which C2 is more advanced than C1. For each level, CEFR describes what a learner is supposed to be able to do in reading, listening, speaking and writing (Little, 2006). The six CEFR reference levels imply that language learners of all levels can be clustered into six groups because there are distinctions in language abilities that exist to separate them as such. Meaning that learners in A1 form a homogenous group relative to learners in B1 or in C2 because distinctions exist among what learner are able to do (or conversely not able to do) with respect to other levels (e.g., understand a wide range of demanding, longer texts, and recognize implicit meaning, etc.).

Classifying learners into proficiency levels involves linking test scores into levels of target language ability (Kaftandjieva, 2010). The challenge in placing students in this scenario is that proficiency placement tests scores are numeric, whereas CEFR, for example, are categories based on qualitative descriptions of learner abilities. This means that based on test scores, learners need to be categorized into appropriate levels of proficiency based on quantitative cut-off scores (Kim et al., 2017). While standardized testing such as TOEFL[®] or IELTS[®] have mapped test scores to CEFR based on descriptors of levels with careful analysis of test content, quantitative analysis that reproduces these categories using examinees' actual scores have been lacking in the literature. That is, whether clusters of levels do exist in data has not been adequately addressed. This study attempts to fill this gap by investigating the dataset of TOEFL[®] scores from the ICNALE (International Corpus Network of Asian Learners of English) learner corpus to recreate the CEFR levels. In order to explore the possibility of learner clusters within the distribution of test scores, CEFR levels and their mapping to TOEFL[®] scores were analyzed for maximum separation based on total scores. This was done by finding the maximum differences in the distribution of scores through a univariate clustering of test scores using the Kernel Density Estimation method.

2. Review of Literature

Every year across the globe, thousands of non-native speakers of English enroll in classes either abroad or in their native countries in both private and public language institutions in order to improve their English proficiency. Before the classes can begin, however, language learners must be assessed on their current level of proficiency, then placed into appropriate levels of ESL classes (Callahan et al., 2010). There is an obvious motivation that justifies the practice of conducting proficiency placement. Learners need to be placed into appropriate classes designed to meet their educational needs that challenge their level of understanding and help them advance towards better proficiency and mastery of the target language. Certainly, learners

need to be placed into their appropriate level, and they also need to be placed with peers who share similar proficiency profiles with their classmates (Maros et al., 2012). When a classroom has a wide range of abilities and background knowledge it poses a pedagogical challenge for the instructor because the materials presented may be either too challenging or too easy for portions of the class. Furthermore, language classes are intended to be interactive as students often engage in pair-work, group talks and class discussion that encourage target language use (Long, 1996). Therefore, when placing students into classes, each grouping aims to maximize homogeneity in learner proficiency, even though in reality no two learners are exactly identical in their foreign language competence, and individual differences are expected to be found for all learners (Dörnyei & Skehan, 2003; Gardner & Lambert, 1959, 1965). The challenge to placement is not only to identify learners based on their language proficiency but also to minimize intra-group differences while maximizing inter-group differences in language abilities (Maros et al., 2012).

In this regard, learners' foreign language proficiency is generally measured through numeric scoring on a proficiency test. Learners are then placed in categorial levels indicative of their proficiency (Nallaya, 2012). This procedure may seem straightforward, but in practice it poses two problems for language institutions tasked with placing students: (1) how to map placement test scores into proficiency levels and (2) how many levels to map them into. The latter problem can be resolved by adopting previously defined frameworks such as CEFR reference levels in conjunction with the governing bodies' limiting resources. That is, *inter alia* each institution's limitations on the number of instructors, classroom space, and operating budget. The former problem about mapping numeric scores to categorial proficiency levels often requires establishing cut scores that demarcate score ranges for each level. The aforementioned problems described are referred to as the "cut-score conundrum" according to Cizek (1993) because absolute and true cut scores do not exist and learners do not come with easy to identify labels of their proficiency.

The process of establishing cut scores on a given test is known as *standard setting* in the field of educational testing and measurement (Cizek, 1993; Jaeger, 1989). In standard setting, scholars, instructors and/or test developers agree that some levels of subjectivity and controversy are bound to occur even among experts tasked with establishing what specific number constitutes an accurate separation of language competence (Cizek, 1993; Cizek & Bunch, 2007). In reality no single number exists that can perfectly separate learners into their appropriate levels. Nevertheless, the goal of standard setting is to establish cut scores that are principled and empirically justified. One prime example of a guideline that maps an English proficiency test score to CEFR levels is ETS's (Educational Testing Services) conversion of TOEFL® test scores to CEFR levels of C1, B2, B1, and A2. Table 1 below shows how TOEFL® scores can be referenced to reflect the examinees' CEFR levels according to ETS.

Table 1.
CEFR Level to TOEFL® score

CEFR level	Total (0–120)
C1 or above	95
B2	72
B1	42
A2	n/a

(Papageorgiou et al., 2015)

Note. Only four CEFR levels are mapped to TOEFL® score because TOEFL® was not designed *a priori* to map to the CEFR; therefore, not all six levels of CEFR were found to be identifiable (Tannenbaum & Wylie, 2008).

It is important to note that the TOEFL® test and its scores were not designed to create differentiation around cut scores. According to ETS research report, "...the TOEFL® iBT test does not target any one specific level of language proficiency, but rather is designed to assess a range of proficiency levels through performance on a variety of assessment tasks..." (Papageorgiou et al., 2015, p.9) and that learners' proficiency is considered to be a continuous construct that spans a wide spectrum of learner abilities comprised of multiple modalities. The construction of cut scores in Table 1 by ETS involved a long process of reviews, analysis, discussion, and research by panelists tasked with creating a link between TOEFL® and CEFR. A panel of twenty-three experts who were familiar with the TOEFL® test and CEFR reference levels were selected in the standard setting process. An iterated modified Angoff approach was implemented for the listening and the reading section of the TOEFL® test. The Angoff method (1971) proposes expert decision-makers to review each test item and estimate the proportion of the target subpopulation of examinees who would answer the item correctly. The subpopulation of interest to be considered would have the minimally competent ability to delimit the cut-off borderline between the different subpopulations. In practice, ETS asked its select panelists to estimate the probability that a barely-passing examinee would answer the item correctly. The sum of each panelist's probability judgment was used to create the recommended cut-off score. A similar performance-sample approach was implemented for the speaking and the writing sections. Panelists were asked to identify the performance profile score for just-qualified candidates based on the review of the scoring rubric (for details refer to Tannenbaum & Wylie, 2008). In short, the modified Angoff method and the performance-sample approach both require experts to judge the most probable outcome on each item or task by borderline *hypothetical* examinees.

According to Jaeger (1989), these types of standard-setting methods are characterized as "test-centered" methods because the cut scores are established based on the item/task analysis of the tests. Cut-scores based on the test's ability to discriminate examinee ability accounts for the majority of standard-setting methods used (Cizek & Bunch, 2007) in most testing scenarios. In practice, test-centered methods of standard-setting require analyzing the test items to determine the most probable cut-score for the proficiency level of interest. For

example, this means that the establishment of cut scores for the binary case of masters versus non-masters are based on how well a non-master/master level examinee might do on an item to item basis. Thus, a test-centered method requires an expert judgment of language abilities of examinees as measured by the test items. In contrast, a second class of method known as “examinee-centered” uses information about *real* examinees and their actual performance— as opposed to hypothetical examinees— to make decisions about where the cut-off scores should be. For example, the Contrasting Groups method (Berk, 1976) plots the total score of all subpopulations of interest then finds points on the score scale that would maximize the differences between the subpopulations or the proficiency levels in the case of placement tests. In the binary case of masters vs. non-masters, one example of examinee-centered method looks for patterns that are indicative of distinct clusters in the overall distribution of test scores. Thus, in examinee-centered methods, judgements that separate examinee into distinct groups are bottom-up and data driven. In actual practice, finding the points of separation between distinct ability groups most often involve using logistic regression for dichotomous sub-groups (e.g., decision between admission acceptance versus rejection) or multinomial logistic regression for multilevel groups, also known as polytomous, logistic regression model (for in-depth example see Bersabé & Rivas, 2010). The cut score is simply defined as the score at which the examinee is likely to be in the category by solving for the value of the score using the model parameters obtained from the regression model. For example, in the case of a single cut score, logistic regression can be used to determine the score at which the probability of membership is .50 by solving the regression equation with the parameters obtained from the data (Bersabé & Rivas, 2010; Cizek & Bunch, 2007). This requires a meticulous standard-setting of each examinee being carefully categorized into proficiency levels by some other standards, followed by cut score determination *post-facto* test administration. The limitation, however, of using regression or similar methods to determine cut scores is that it still requires experts’ *a priori* knowledge of the examinees’ proficiency membership. Thus, there is a danger of committing a tautological fallacy in that in order to determine the cut score for establishing proficiency levels, one must begin with judgments about their proficiency to fit the model. In other words, regression is a type of supervised learning algorithm when classifying learners into proficiency levels is more akin to unsupervised learning problem in reality (James et al., 2013).

What is proposed in this study is a true bottom-up approach to standard-setting, and one that does not require knowing the proficiency level labels of each participant— a more likely scenario in a typical placement setting. Instead of using regression to determine cut scores, an alternative in contrasting groups method is to use kernel density estimation (KDE). For a univariate, one-dimensional data such as total test score, kernel density estimation is an unsupervised learning procedure appropriate for classifying participants (Hastie et al., 2016; Kaufman & Rousseau, 1990).

Let (x_1, x_2, \dots, x_n) be a univariate (i.i.d.) drawn from a probability density $f(x)$. In estimating the shape of the distribution, the kernel density estimator is the following:

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (1)$$

In the above Equation (1), k is the non-negative, symmetric kernel function and h is the smoothing, bandwidth parameter. Simply put, each data point is centered on the kernel function then summed together on all data points to create the kernel density estimate (James et al., 2013). This transforms the shape of the distribution by binning the scores to create clusters based on separations in the score represented by the kernel density estimation. If KDE is to produce an efficient and effective method to separate the learners, the distribution is expected to be multimodal as each proficient level is represented by its own peak in the overall distribution of total scores. Therefore, the challenge in KDE is finding the optimal choice of kernel-bandwidth to maximize the differences in peaks found in the data. One effective method is to do a grid search cross-validation (Matioli et al., 2018; Salgado-Ugarte & Pérez-Hernández, 2003). The basic premise is that the data is split into K subsets, followed by training the KDE model with $K-1$ data-points and evaluating the performance on the holdout data. The hyperparameter (bandwidth) that gives the best performance on average across all K test folds is chosen as the bandwidth for binning the data. The evaluation of performance on the test set is based on log-likelihood of the test data in the estimated kernel density. The statistical details are beyond the scope of this paper but many sources are available regarding the details of this process (e.g., Silverman, 1986; Wand & Jones, 1995) as well as statistical packages and models available in R and Python using Scikit-Learn.

Based on the review of how TOEFL[®] test scores are mapped to CEFR levels, standard setting of cut scores through the examination of real examinee score distribution is proposed in the current study. The research questions were motivated by a lack of bottom-up approach to standard setting that is also examinee-centered in the previous research in language testing.

2.1. Research questions

There were three research questions in the current study:

- 1) Does the KDE distribution of scores on the TOEFL[®] test demonstrate that learners can be clustered based on similarities in their total score?
- 2) Does the KDE distribution of scores on the TOEFL[®] test confirm the use of cut scores established by ETS to divide learners into four levels?
- 3) What are the differences (if any) between the ETS cut scores and KDE generated cut scores?

3. Method

3.1. Data

The ICNALE (International Corpus Network of Asian Learners of English) is an international English learner corpus developed and maintained by Shin Ishikawa at the Kobe University in Japan. The ICNALE, which has become one of the largest learner corpora publicly available, houses more than 10,000 speeches (both monologues and dialogues) as well as essays (edited and original) produced by college and graduate students. The Asian countries represented by ICNALE include China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore/ Malaysia, Taiwan, and Thailand. There are four sub corpora which comprise the ICNALE: (1) Spoken Monologue, (2) Spoken Dialogue, (3) Written Essays, and (4) Edited Essays. For the purposes of this study, three sub-corpora from ICNALE were used: Spoken

Monologue (Ishikawa, 2014), Written Essays (Ishikawa, 2013), and Edited Essays ((Ishikawa, 2018). When these three modules were combined, the total sample size was 939 English learners. In addition to the corpus of speeches and essays, the ICNALE has data on students' proficiency test scores such as TOEFL[®] iBT (internet-based), paper-based TOEFL[®], IELTS, and TOEIC. TOEFL[®] iBT scores from ICNALE was used to address the research questions (Table 2). Not all students listed in the corpora had TOEFL[®] iBT scores: some had IELTS, others had paper-based TOEFL[®] and the rest had TOEIC scores. In these instances, their equivalent scores in TOEFL[®] iBT were calculated according to the ETS guidelines that link test scores to TOEFL[®] iBT (ETS, 2005 & 2010). Based on Table 1, the breakdown by CEFR level was the following: A2 (N=62); B1 (N=345); B2 (N=245); C1 (N=287) for a total N=939.

Table 2.

Descriptive Summary of TOEFL[®] Score (N=939)

Variable Name	Description of Variables	Values	Mean	Std
TOEFL [®]	Test of English as a Foreign Language	0 – 120	76.60	23.07

3.2. Method of analysis

Python 3 and R were used for data visualization and data analysis. For the first research question, there were two methods of analysis. First, the TOEFL[®] scores of 939 examinees were visualized with a histogram in R. Next, the distribution of total scores were analyzed using Kernel Density Estimation (KDE) method in Python's machine learning library 'Scikit-Learn'. Different bandwidths were used to create a smooth estimate over the distribution of TOEFL[®] scores. Finally, the grid search cross-validation method was used to discover the optimum bandwidth. The clusters of learners were determined based on the local minima of the KDE distribution and the results were compared to the ETS cut scores via confusion matrix.

4. Results and Discussion

Before applying KDE to the data, a histogram was produced in order to provide a quick visualization of the data. The histogram of TOEFL[®] scores was generated using Sturges' rule for binning (Figure 1). Sturges' rule is a well-known and most commonly applied rule defined as $w = r / (1 + \log_2(n))$ in which r is the range of values within the data set, and n is the number of elements within the data. The result is the ideal bandwidth w for the histogram (Wand, 1997). The reason for creating a histogram with Sturges rule is that this is often the default set value for generating basic histogram in statistical software such as R. This provides perhaps the most basic, quick and simple method to visualize the distribution of the variable of interest without delving into more complicated statistical analysis.

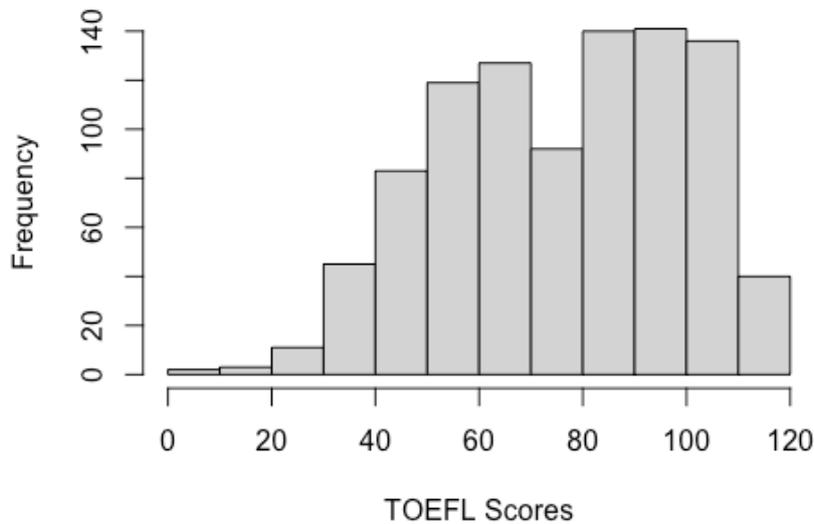


Figure 1. Histogram of TOEFL[®] Scores

Figure 1 above shows the frequency distribution of TOEFL[®] scores. It appears that there are at least two peaks around scores of 60 and 90, which could be interpreted as clusters of proficiency levels. However, visualizing the most optimal shape of the distribution that does not “over-smooth” the curve or capture too much noise in the data is the most crucial step in discerning whether multiple peaks or clusters exist in the data (James et al., 2013). Thus, simply looking at the histogram of total scores is inappropriate to draw meaningful conclusions about the existence of cut scores. For this reason, the Kernel Density Estimation method was implemented to create an optimal shape of the score distribution, most appropriate for examinee-centered standard-setting that maximizes and accentuates the separation of peaks in the distribution.

Figure 2 below shows the Kernel Density Estimation of the TOEFL[®] scores using optimal bandwidth estimate with Gaussian kernel (Hastie et al., 2016). Gaussian kernel was chosen to model the assumption that standardized tests such as the TOEFL[®] scores are normally distributed (Allen & Yen, 2002). After grid search optimization (cross validation), the optimal bandwidth was found to be 1.95. The local minimums were labeled with dotted vertical lines.

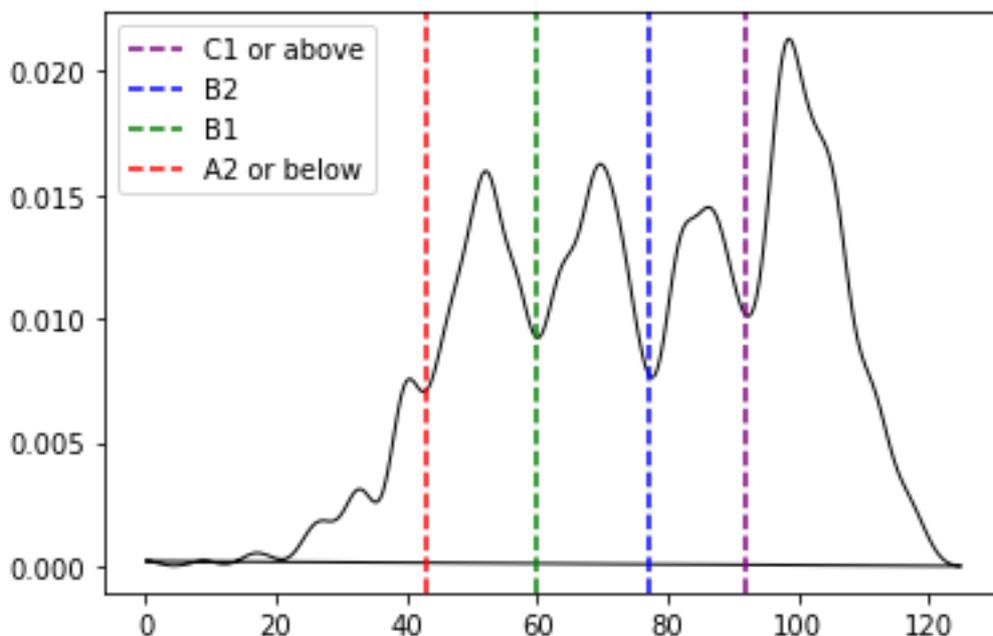


Figure 2. Optimal Kernel Density Estimation of TOEFL® scores

Based on KDE, four local minimums were found at scores of 92, 77, 60, and 43. According to these points the cut scores can be divided into (1) >92; (2) <=92 and >= 77; (3) <77 and >=43; (4) <42. In other words, the highest cluster was found to be at the border of 92 or above, followed by a cluster in the range of scores between 77 to 91, followed by a cluster in the range of scores between 60 to 76, and lastly, a cluster of students in the range of scores below 43. A side by side comparison of cut scores provided by ETS and KDE is shown in Table 3 below.

Table 3.
TOEFL® cut scores by ETS and KDE Comparison

CEFR level	ETS	KDE Based
C1 or above	95	92
B2	72	77
B1	42	60
A2	n/a	n/a or 42

According to Table 3, the breakdown by CEFR levels according to KDE was the following: A2 (N=261); B1 (N=199); B2 (N=168); C1 (N=311), compared with the breakdown by CEFR level based on the ETS’s standards: A2 (N=62); B1 (N=345); B2 (N=245); C1 (N=287). By applying the cut score ranges provided by ETS and ones generated through KDE on the dataset, a confusion matrix was constructed (Table 4). Comparison between the two methods on their classification of four CEFR levels can be seen below.

Table 4.

Confusion Matrix Comparing ETS vs KDE

	KDE C1 or above	KDE B2	KDE B1	KDE A2 or below	Total ETS
ETS C1 or above	287	0	0	0	287
ETS B2	24	168	53	0	245
ETS B1	0	0	146	199	345
ETS A2 or below	0	0	0	62	62
Total KDE	311	168	199	261	939

The confusion matrix shows that for C1 or above there was a general agreement between ETS and KDE, except 24 examinees identified as B2 by ETS which were labeled as C1 or above by KDE. For B2, in addition to 168 identified by KDE, ETS added 24 (identified as C1 or above by KDE) and 53 (identified as B1 by KDE). The largest discrepancy was found for B1 and A2 categories in which 199 from KDE method labeled as A2 were identified as B1 by ETS. This was equally reflected in the drastic difference in the cut score of 42 versus 60 found in Table 3 above.

Regarding the research question one, the kernel density estimation on the univariate data of *TOEFL® iBT* total score produced results that seem to indicate that four groups of learners per CEFR were reasonable. Figure 2 showed that there were four discernable peaks in the overall distribution of scores. Based on the Contrasting Groups method, the use of peaks as boundaries qualifies as maximizing differences among subpopulations. In fact, the four peaks may be an indication that the overall distribution of total score is a mixture of four subpopulations of learner abilities, each with its own mean and standard deviation. This is what would be expected for clusters of learners comprised of the same proficiency profiles yet distinct from others of different proficiency levels (Bond & Fox, 2015). Therefore, the multimodal distribution of scores supports the notion of grouping learners not simply based on cut scores of measured abilities but also on learners' similarity within each distribution. In short, there was enough evidence from the examinee-centered approach to validate mapping *TOEFL®* scores to four clusters of learners that differed on their level of proficiency.

The second research question asked whether KDE distribution of scores on the *TOEFL®* test confirms the use of cut scores established by ETS. Table 3 seems to show that the cut scores established by ETS fit the KDE distributional peaks for CEFR level C1 or above and CEFR level B2. The last research question was addressed by the confusion matrix in Table 4, which showed that there was a sharp discrepancy between ETS and KDE on the boundaries of B1 and subsequently A2. KDE of Figure 2 indicates that the local minimum that separated its two adjacent peaks is at 60, not 42. Based on this, the next cluster of learner proficiency that

is one level below B2 was found to be 60. This does not invalidate ETS's demarcation at 42, as the score of 42 may very well be the most appropriate descriptor of B1 per CEFR and concurrently established based on the panel of experts at ETS (Little, 2006; Papageorgiou et al., 2015; Tannenbaum & Wylie, 2008). Rather, the real issue is that the score of 42 was much closer to KDE's last local minimum at 43. One interpretation of this is that the English language abilities of two peak-groups between scores of 43 and 77 may not be distinguishable in terms of CEFR's characterization of reference levels (B1), but they were *distinguishable* nonetheless as a separate cluster of learners with lower proficiency than the group with a peak between 92 and 77. What could be the cause of difference found in cut scores for B1 warrants further investigation and discussion.

First, it is important to be reminded that there is no single absolute cut-score free of controversy. In fact, a single cut-score for all non-native speakers of English under any and every context would be highly inappropriate because standard-setting process is a meticulous and principled procedure that should be context dependent. In this regard, it is worth repeating Cizek and Bunch's (2007) comment about the need to consider the population demographics. "...any judgmental standard-setting process necessarily requires participants to bring to bear information on both test content and the population on which the performance standards will be set." (Cizek & Bunch 2007, p. 105).

When ETS first linked *TOEFL*® *iBT* test scores onto the CEFR, a panel of experts comprised of 23 members were tasked with developing the cut-score boundaries that corresponded to CEFR level descriptors (Tannenbaum & Wylie, 2008). It is crucial to note that this panel was represented by 16 countries that were mostly from Europe with the exception of two from Turkey and one from United Arab Emirates. Likewise, in 2015, when the same cut scores were revised by Papageorgiou and colleagues (2015), the revisions were made in part based on feedback from university programs in Europe. In their research report, Papageorgiou et al. (2015) investigated the reasonableness of the revised cut scores by reviewing admissions information of 100 universities in the United States; 30 in the United Kingdom; 15 in Canada; and 10 in Australia. What this demonstrates is that the examinee demographics in which ETS developed its cut score is different from the examinee demographic of the current study: ICNALE (International Corpus Network of Asian Learners of English). The differences between ETS and KDE's derived cut score were not only methodological in terms of test-centered versus examinee-centered, they were also different in terms of the population on which the performance standards were set. One may object that after all Common *European Framework of Reference* was developed to provide a common basis for language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe (Little, 2006). However, there is nothing in the CEFR descriptors that would belie a framework that would be equally valid across all nationalities. Besides, *TOEFL*® *iBT* test is not restricted to one continent but it is internationally accepted across multiple nations across the globe. Therefore, by equating *TOEFL*® to CEFR, there was no reason to believe that CEFR would be restricted to European nationalities only. The critical issue is that as previously mentioned, standard-setting process necessarily requires consideration of the population and not just the test content, and consequently different population of examinees may warrant different cut scores as a result.

5. Conclusion

In this study, kernel density estimation of total TOEFL[®] iBT scores from ICNALE has shown that learners can be grouped into four distinct clusters based on proficiency as measured by the TOEFL[®] test. In addition, the boundaries of these learner-clusters were shown to be mostly comparable to cut-score guidelines provided by ETS for the purpose of mapping TOEFL[®] scores to CEFR levels C1, B2, B1, and A2. The discrepancy between ETS and KDE was found for B1. The results demonstrated that an examinee-centered method of standard-setting such as contrasting group method via KDE has the potential to reproduce a test-centered method like the modified Angoff method implemented by ETS. However, the discrepancy between the two methods points toward a need to combine both approaches to fine-tune the standard-setting process. The distributional characteristics of Asian learners may be different from European students on which ETS has based its mapping (Bolton, 2008). If this is the case, stakeholders and institutional governing bodies that determine cut scores must incorporate both test-centered methods and examinee-centered methods so that both test content and the population on which the performance standards will be set are taken into consideration. This is especially germane in high-stakes language tests such as TOEFL[®] iBT. The call for considering demographic information of test takers into account should be the norm of stand-setting rather than the exception (Rezaeian et al., 2020). Finally, there are important implications for future practice of standard-setting in language assessment based on the findings of this study. First, the use of unidimensional cluster analysis, such as KDE can be employed in addition to the top-down method of establishing cut-scores. Second, the data driven approach to finding natural group clusters may provide a context-dependent method of standard setting in which the demographic information of the examinees are taken into consideration by letting the data “speak for itself”. Imposing cut scores without understanding the distributional characteristics of the test score could lead to a non-optimal separation of learners that does not categorize them into their most natural proficiency groups.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC American Council on Education.
- Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45(2), 4–9.
- Bersabé, R., & Rivas, T. (2010). A general equation to obtain multiple cut-off scores on a test from multinomial logistic regression. *The Spanish journal of psychology*, 13(1), 494–502. <https://doi.org/10.1017/s1138741600004042>

- Bolton, K. (2008). English in Asia, Asian Englishes, and the issue of proficiency. *English Today*, 24(2), 3-12.
doi:<http://dx.doi.org.ezproxy.cul.columbia.edu/10.1017/S026607840800014X>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in human sciences (3rd ed)*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Callahan, R., Wilkinson, L., & Muller, C. (2010). Academic achievement and course taking among language minority youth in U.S. schools: Effects of ESL placement. *Educ Eval Policy Anal*, 32(1), 84–117.
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30(2), 93–106.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J., Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Blackwell.
- Educational Testing Service. (2005). *TOEFL® Internet-based Test Score Comparison Tables*. Retrieved December 18, 2020 from https://www.kuleuven.be/english/admissions/lang/TOEFL_iBT_Score_Comparison_Tables.pdf
- Educational Testing Service. (2010). *Linking TOEFL® iBT Scores to IELTS Scores. A Research Report*. Retrieved December 18, 2020 from https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf
- Gardner, R. C., & Lambert, W. E. (1959). Motivational variables in second language acquisition. *Canadian Journal of Psychology*, 13(4), 266–272.
- Gardner, R. C., & Lambert, W.E. (1965). Language aptitude, intelligence, and second-language achievement. *Journal of Education Psychology* 56(4), 191–199.
- Hastie T., Tibshirani R. & Friedman J. (2016). *The elements of statistical learning: data mining, inference, and prediction*, 2nd Ed. Springer.
- Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp. 91-118). Kobe, Japan: Kobe University.
- Ishikawa, S. (2014). Design of the ICNALE Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2* (pp. 63-76). Kobe, Japan: Kobe University.
- Ishikawa, S. (2018). The ICNALE edited essays; A dataset for analysis of L2 English learner essays based on a new integrative viewpoint. *English Corpus Studies*, 25, 117–130.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York Macmillan.
- James, G., Witten, D., Hastie T., & Tibshirani R. (2013). *An introduction to statistical learning: with applications in R*. Springer.
- 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: Cito.

- Kaufman, L., & Rousseau, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley.
- Kim, B., Kim, J.M., & Yi, G. (2017). Analysis of clustering evaluation considering features of item response data using data mining technique for setting cut-off scores. *Symmetry* 9(5), 1–17.
- Little, D. (2006). The Common European Framework of Reference for Languages: Contents, purpose, origin, reception and impact. *Language Teaching*, 39(3), 167-190.
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.
- Maros, M., Stapa, S., & Yasin, M. (2012). English language proficiency levels and needs of international postgraduate students: implications and recommendations. *Asian Social Science*. 8(13), 181–187.
- Matioli, L. C., Santos, S. R., Kleina, M., & Leite, E. A. (2018). A new algorithm for clustering based on kernel density estimation. *Journal of Applied Statistics*, 45(2), 347–366. <https://doi-org.ezproxy.cul.columbia.edu/10.1080/02664763.2016.1277191>
- Nallaya, S. (2012). The measurement of change in English language proficiency. *Issues in Educational Research*, 22(2), 149–168.
- Papageorgiou, S., Tannenbaum, R., Bridgeman, B., & Cho, Y. (2015). The association between TOEFL® iBT test scores and the Common European Framework of Reference (CEFR) levels. *Research Memorandum*, ETS RM-15-06.
- Rezaeian, M., Seyyedrezaei, S.H., Barani, G., & Seyyedrezaei, Z.S. (2020). Construction and validation of educational, social and psychological consequences questionnaires of EPT as a high-stakes test. *International Journal of Language Testing*, 10(2), 33–70.
- Salgado-Ugarte, I. H., & Pérez-Hernández, M. A. (2003). Exploring the use of variable bandwidth kernel density estimators, *The Stata Journal*, 3(2), 133–147.
- Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall.
- 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* (TOEFL® iBT Research Report RR-08-34). Princeton, NJ: Educational Testing Service.
- Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall.
- Wand, M.P. (1997). Data-based choice of histogram bin width. *The American Statistician*, 51(1), 59–60.

APPENDIX A

CEFR Reference levels and their descriptions

Retrieved from <https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>

PROFICIENT USER	C2	Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.
INDEPENDENT USER	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
BASIC USER	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.