Measuring the Power of Learning.™

Research Memorandum
ETS RM–15-06

The Association Between *TOEFL iBT®* Test Scores and the Common European Framework of Reference (CEFR) Levels

Spiros Papageorgiou

Richard J. Tannenbaum

Brent Bridgeman

Yeonsuk Cho

August 2015

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

The Association Between *TOEFL iBT*® Test Scores and the Common European Framework of Reference (CEFR) Levels

Spiros Papageorgiou, Richard J. Tannenbaum, Brent Bridgeman, and Yeonsuk Cho
Educational Testing Service, Princeton, New Jersey

August 2015

Corresponding author: Spiros Papageorgiou, E-mail: spapageorgiou@ets.org

Suggested citation: Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

Action Editor: Donald Powers

Reviewers: Jonathan Schmidgall and Michael Kane

Abstract

The Common European Framework of Reference (CEFR), published by the Council of Europe (2001) is arguably one of the most influential language frameworks in the field of second language teaching and assessment, articulating a progression of language proficiency through a number of levels. Tannenbaum & Wylie (2008) mapped the *TOEFL iBT*® test scores onto the CEFR levels to help test users and decision makers interpret TOEFL iBT test scores in terms of the CEFR levels. Based on the feedback of subsequent users and decisions makers, Educational Testing Service (ETS) revised the CEFR cut scores (i.e., minimum test scores required for each CEFR level) in 2014. In this research memorandum, we present the rationale for the revision of the CEFR cut scores and offer validity evidence that the revised cut scores (a) are reasonable and (b) do not negatively impact the quality of admissions decisions.

Key words: CEFR, cut scores, language proficiency levels, score interpretation, *TOEFL iBT*®

Current conceptualizations of validity and the process of validation place emphasis on the interpretation of a test score, its use, and the impact of that use (Bachman, 2005; Bachman & Palmer, 2010; Kane, 2006, 2013). Scores on language tests for speakers of English as a second/foreign language (ESL/EFL) are often used to classify test takers into different categories or levels of proficiency. In academic contexts, for example, *TOEFL iBT*® test scores are used by universities employing English as the primary mode of instruction to determine whether prospective ESL students have sufficient English-language skills in order to be admitted (Chapelle, Enright, & Jamieson, 2008; Cho & Bridgeman, 2012). As Tannenbaum and Cho (2014) noted, these types of decisions are criterion based, in that a defined level of language proficiency should be met.

However, a test score by itself does not indicate if the criterion has been met. One way to relate test scores to criteria is to map (i.e., associate or link) test scores with descriptions of levels of language proficiency (Tannenbaum & Cho, 2014). The Common European Framework of Reference (CEFR; Council of Europe, 2001) is probably the most influential language frameworks in the field of second language teaching and assessment articulating a progression of language proficiency through six main levels.

It is not easy to establish whether and to what extent admission decisions into higher education are made in relation to the CEFR levels because no uniform policy exists across institutions or educational authorities. In their study, Carlsen and Deygers (2014) argued that B2 level is the most common requirement for admissions into European universities. For example, at the time of producing this research memorandum, the UK government required evidence of English-language proficiency at B2 level for students applying for a Tier 4 student visa to pursue an academic degree in the country.[1] Kantarcioglu, Thomas, O'Dwyer, and O'Sullivan (2010) also reported the same CEFR level requirement (B2) for students in an English-medium university in Turkey. However, in North America and other parts of the world outside Europe, where TOEFL iBT test scores are used to inform admission decisions, reference to the CEFR to set score requirements seems to be much less common, with universities, for example, setting their own, context-specific requirements, which can vary a lot from institution to institution (see, for example, Ling, Wolf, Cho, & Wang, 2014).

The CEFR can be a useful tool for informing decisions about levels of English-language proficiency. However, it should be kept in mind that the CEFR was designed as a generic

reference document (as its title clearly indicates) so that it can be applied in a variety of contexts (Milanovic & Weir, 2010). Although several of its language proficiency descriptors appear to be relevant to the use of academic language (see North, 2014, pp. 62−65), admissions decisions are likely to be based on a variety of factors that go beyond a generic description of language proficiency such as the one found in the CEFR descriptors. This practice of making decisions for academic admission is because setting cut scores is a context-specific, value-driven process (Kane, 2001; Tannenbaum & Katz, 2013), as two recent studies demonstrate with regard to the use of cut scores of English-language proficiency tests (Ling et al., 2014; Papageorgiou & Cho, 2014). For these reasons, users of the TOEFL iBT test are encouraged to set their own score requirements in order to better serve their local needs (Educational Testing Service [ETS], 2005). In the process of setting requirements, users are also encouraged to consult empirically derived performance descriptors that provide additional evidence about the expected English proficiency of test takers at differing TOEFL iBT test score ranges (see, for example, ETS, 2014; Garcia Gomez, Noah, Schedl, Wright, & Yolkut 2007). For test users and decision makers who wish to interpret TOEFL iBT test scores in terms of the CEFR levels in order to inform their decisions, Tannenbaum and Wylie (2008) conducted a study that mapped TOEFL iBT test scores to these levels.

Since the time of the mapping study (Tannenbaum & Wylie, 2008), ETS has been monitoring the needs of the above test users and decision makers and how they use the proposed CEFR cut scores (i.e., minimum test scores required for each CEFR level) to inform their admissions requirements in relation to English-language proficiency. Recall, to our knowledge, many university programs in Europe consider B2 to represent the constellation of English skills likely sufficient to cope with university instruction conducted in English—and hence, to be sufficient for use as one criterion for admissions. Feedback from these users and decision makers, mostly universities that use CEFR levels to define admissions standards in the UK and other European countries, suggested that the TOEFL iBT test score mapping results to the CEFR levels might have been too rigorous, resulting in higher test scores than perhaps needed to reflect the English skills described by the B2 level (and other levels). Moreover, as ETS assessment developers and score users obtained a better understanding of the CEFR scales and their descriptors in the intended target language use (TLU) domain (Bachman & Palmer, 2010) for the TOEFL iBT test (i.e., postsecondary academic), it was reasonable to reconsider the relationship

between test scores and the CEFR levels (see relevant discussion in Taylor, 2004). As a result of considering all the above information, and as suggested in the standard-setting literature (e.g., Geisinger & McCormick, 2010), a revised set of CEFR cut scores for the TOEFL iBT test was proposed. The rationale behind the revision is presented in this report.

Although the revised cut scores reflected, in part, the feedback received from decision makers at universities that use CEFR levels to define admissions standards (mostly universities in the UK and other European countries), the reasonableness of these revised cut scores and their impact on admissions needed to be investigated. Such investigation is the focus of the work documented in subsequent sections in this report. Following an argument-based approach (Kane, 2006, 2013), we aim, through the use of external, nonassessment criteria (Kane, 2001), to provide evidence supporting two claims related to the inferences that can be made on the basis of TOEFL iBT test scores:

- Claim 1 (reasonableness of the cut scores): The revised CEFR cut scores are reasonable for making decisions about admission into higher education.
- Claim 2 (impact of the cut scores): The revised CEFR cut scores do not negatively impact admissions decisions due to classification errors.

Before discussing the analyses providing support to the above claims, we first present a brief overview of the CEFR and the process of mapping test scores to its levels.

Mapping Test Scores to the Common European Framework of Reference (CEFR)

The CEFR is one of several publications of the Council of Europe, which have been influential in second language teaching since the 1970s (Van Ek & Trim, 1991, 1998, 2001; Wilkins, 1976). According to the Council of Europe (2001), a common framework for learning, teaching, and assessment is desirable to

- promote and facilitate cooperation among educational institutions in different countries;
- provide a sound basis for the mutual recognition of language qualifications; and
- assist learners, teachers, course designers, examining bodies and educational administrators in situating and coordinating their efforts. (p. 5)

Although the CEFR contains rich information about the language learning process and teaching as well as assessment in nine chapters and four appendices, its language proficiency scales[2] are arguably the best known part of the 2001 volume (Little, 2006). The CEFR scales and descriptors were primarily developed during a large research project in Switzerland (North, 2000; North & Schneider, 1998). The proficiency scales of the CEFR have gained popularity because they offer a comprehensive description of the objectives that learners can expect to achieve at different levels of language proficiency. They describe language activities and competences at six main levels: A1 (the lowest) through A2, B1, B2, C1, and C2 (the highest). The scales comprise statements called "descriptors," which are always phrased positively, as they are intended to motivate learners by describing what they can do when they use the language, rather than what they cannot do (Council of Europe, 2001, p. 205).

The CEFR proficiency scales provide a convenient structure for thinking about and communicating a progression of language proficiency and for considering where people stand in relation to that progression. Therefore, mapping language test scores onto the CEFR levels is a useful way to assign practical meaning to those scores. For example, if a score of at least 16 on a speaking test were associated with the CEFR B1 level, that would suggest that test takers with at least a 16 are, among other things, likely to be able to "enter unprepared into a conversation on topics that are familiar" and "briefly give reasons and explanations for opinions and plans" (Council of Europe, 2001, p. 26). To further help test providers add meaning to their test scores in relation to the CEFR levels, the Council of Europe (2009) published a manual offering a recommended set of procedures for aligning both test content and test scores with the CEFR levels. The CEFR has had a wide impact in Europe, and its six main levels "have become a common currency in language education, and curricula, syllabuses, textbooks, teacher training courses, not only examinations, claim to be related to the CEFR" (Alderson, 2007, p. 660). Applications of the CEFR in these areas are illustrated by several studies presented in three edited volumes (Byram & Parmenter, 2012; Figueras, & Noijons, 2009; Martyniuk, 2010) and also North (2014).

A number of studies and research projects such as the DIALANG project (Alderson 2005; Alderson & Huhta, 2005; Kaftandjieva & Takala, 2002) have shown that the hierarchy of the CEFR language proficiency descriptors can be consistently replicated in a range of contexts, thus offering validity evidence for the use of those descriptors and the scales they belong to

across a variety of contexts. However, the CEFR is neither a static tool nor a prescription to be followed with one singularly correct interpretation or application for designing test content or interpreting test scores. In fact, because the CEFR is intentionally context-free to allow for a variety of applications and its language proficiency descriptors are not specific to a language, researchers note problems when using the CEFR to design test specifications and tasks (Alderson et al., 2006; Hasselgreen, 2012; Weir, 2005). One of the chief architects of the CEFR, Brian North, and his colleagues appropriately reminded us of the intended flexibility of the CEFR: "The CEFR is a concertina-like reference tool that . . . educational professionals can merge or sub-divide, elaborate or summarise, adopt or adapt according to the needs of their context. . . . It is for users to choose activities, competences and proficiency stepping-stones that are appropriate to their local context . . ." (North, Martyniuk, & Panthier, 2010, p. 4). Therefore, a key point to note when using the CEFR in teaching, learning and assessment contexts is, "There is not and never will be an authorised interpretation of the CEFR. That openness is the secret of its success" (North, 2014, p. 5).

The mapping of test scores to the CEFR is essential if the scores are to be interpreted in terms of levels of the CEFR. Mapping is typically accomplished through a standard-setting approach, which is based on expert judgment, and informed by test data, which links test performances to CEFR levels (Council of Europe, 2009; Papageorgiou, 2010; Papageorgiou & Tannenbaum, in press; Tannenbaum & Cho, 2014; Tannenbaum & Katz, 2013).

The process of setting standards is not without criticism, however, in large part due to its inherent subjectivity (North, 2014). Skepticism is also fueled by the acknowledgment in the measurement literature that different standard-setting methods produce somewhat different results (Cizek & Bunch, 2007). However, this is much the same as is expected—but, traditionally readily accepted—that a test taker taking two different forms of the same test will not likely earn the same score (Green, Trimble, & Lewis, 2003). Some ambiguity in test scores and in setting standards is inevitable. Nonetheless, North (2014) argued against use of standard setting in order to establish a relationship between test scores and the CEFR, in particular when the Angoff method (Angoff, 1971), or one of its modified variants, is used. It is worth noting, however, that contrary to North's position, research evidence does support the replicability of Angoff-based results (Tannenbaum & Kannan, 2015). North (2014) proposed item banking and item calibration using item response theory (IRT) as the best alternative to standard setting, especially

when tests are intended to measure more than one CEFR level. It could be argued, however, that it is not clear how this can be done without involving human judgment at least to some extent, that is, the aspect of standard setting for which that particular methodology has been criticized. In fact, a test program would need standard setting at some point either for one or more test forms or for an item bank, with some equating method applied to maintain consistency of the cut scores across test forms. Despite North's criticism, the Council of Europe (2009) recognized the central role of standard setting in relating test scores to CEFR levels because "the crucial point in the process of linking an examination to the CEFR is the establishment of a decision rule to allocate students to one of the CEFR levels on the basis of their performance in the examination" (p. 11).

Mapping TOEFL iBT Test Scores to the Common European Framework of Reference (CEFR)

Among the first published studies mapping English-language test scores to the CEFR was that of Tannenbaum and Wylie (2008). The study employed two standard-setting methods: a modified Angoff approach (Brandon, 2004; Cizek & Bunch, 2007; Plake & Cizek, 2012) for selected-response items and a performance profile approach (Hambleton, Jaeger, Plake, & Mills, 2000; Morgan, 2004; Perie & Thurlow, 2012; Zieky, Perie, & Livingston, 2008) for constructed-response items. The panelists in the study were 23 educators from 16 countries specializing in ESL/EFL. The primary outcome of the study was a set of recommended cut scores—minimum test scores that the educators, on average, judged to be needed to enter different CEFR levels.

While it is useful to map test scores to the CEFR levels, one should not assume that the relationship between a language test and the CEFR is necessarily simple, direct, or established as a one-time event. In such instances, the focus is on recommending the lowest acceptable test score (so-called cut scores) that signals entrance into a level of the CEFR. Moreover, as we discussed previously, it is reasonable to reconsider the relationship between test scores and the CEFR levels in light of decision makers' needs and changing assessment contexts and as test developers and score users obtain a better understanding of the CEFR scales and their descriptors (Taylor, 2004) in relation to the TLU domain. In fact, making adjustments to recommended cut scores to better meet the needs of decision makers is accepted practice (e.g., Geisinger & McCormick, 2010), and following an argument-based approach (Kane, 2006, 2013), evidence

should be collected to support claims about the inferences intended to be made based on these scores.

Since Tannenbaum and Wylie's (2008) study, feedback by users and policy makers using the CEFR levels to inform their admissions decisions suggested that the TOEFL iBT test score mapping results to the CEFR levels might be too conservative in their contexts. Applying a stringent score requirement provides greater confidence that test takers classified into the higher of two adjacent CEFR levels (e.g., B2 instead of B1) deserve that elevated classification; that is, a higher cut score reduces false-positive decisions. On the other hand, a stringent score requirement means that some test takers who merit classification into the higher CEFR level (B2) are, in fact, classified at the lower level (B1); so a higher cut score also increases false-negative decisions. In the context of admissions decisions, a false-negative decision means denying an otherwise qualified student the opportunity to enter a desired program of study, as well as denying the program the benefit of having this student.

The feedback from many institutions relying on the CEFR levels indicated that they believed that lowering the score required, for example, to meet the B2 level (reducing the likelihood of making false-negative admission decision) was a reasonable recommendation, given their experience with incoming students. Even though lowering the requirement would also admit some number of students who were not functioning at a B2 level (a false-positive admission decision), many institutions were more in favor of giving students (test takers) the benefit of the doubt. This policy recognizes that test scores are not perfectly reliable and values erring on the side of supporting test takers—that they likely have the English skills needed to cope with instruction delivered in English. The reasonableness of this decision is also bolstered by the fact that many universities have language support programs for admitted students (such as those described in Ling et al., 2014), which, over time, serve to help rectify false-positive decisions. Consequently, lowering the CEFR cut scores for the TOEFL iBT test was warranted for score users who value the CEFR levels as performance standards.

One relevant source of information that can be used to adjust cut scores in a principled way is the error of measurement associated with the test scores (Cizek & Bunch, 2007; Geisinger & McCormick, 2010). The standard error of measurement quantifies how consistent test scores are—for example, the extent to which test takers would earn the same test score if they took different forms of that test. If a test were perfectly reliable, then the same score would be earned.

No test, however, is perfectly reliable, and so test scores do vary. Typically, there is a 68% chance that a test taker's true score would fall within 1 SEM, and 95% chance that it would fall within 2 SEM (for a detailed discussion of standard error of measurement, see the *TOEFL iBT* Research Insight Series[3]).

ETS reviewed the panel-recommended cut scores on the four test sections presented in Tannenbaum and Wylie (2008) and in 2014 lowered those recommendations by 2 SEM (it should be noted that each test section has a different standard error of measurement). Table 1 presents the revised cut scores, in scaled-score units, corresponding to each of the CEFR levels. Total test scores were computed by summing the revised section-level cut scores. The total minimum score corresponding to B1 is now 42 scaled points, down from 57 scaled points; the total minimum score to enter B2 is now 72 scaled points, down from 87 scaled points; and for C1 it is now 95 scaled points, down from 110 scaled points.

Table 1. Common European Framework of Reference (CEFR) Cut Scores for the TOEFL iBT Test

CEFR level	Reading (0−30)	Listening (0−30)	Speaking (0−30)	Writing (0−30)	Total (0−120)
C2	25				
C1	24	22	25	24	95
B2	18	17	20	17	72
B1	4	9	16	13	42
A2			10	7	
A1			5		

Although these revised cut scores reflect, in part, the feedback received from decision makers at universities that use CEFR levels to define admissions standards (mostly universities in the UK and other European countries), the reasonableness of these revised cut scores and their impact on admissions needed to be investigated. Such investigation is the focus of the work documented in subsequent sections in this research memorandum. In particular, we examined the reasonableness and impact of the revised cut scores in relation to the B2 level, which, as we have discussed, is often acknowledged as the standard for admission into higher education. Prior to presenting the analyses providing support to our claims about the reasonableness and the impact of the revised CEFR cut scores, we discuss in the next section some considerations about the alignment of the content of tests such as the TOEFL iBT test to the language ability described in the CEFR.

Alignment of the Content of the TOEFL iBT Test to the Common European Framework of Reference (CEFR) Levels

A relevant issue regarding CEFR-based score interpretations is the extent to which the content of a test is aligned with the language ability described in the CEFR. Without satisfactory content alignment, as Tannenbaum and Cho (2014) pointed out, there is little justification for conducting a standard-setting study, as the test would lack content-based validity. However, content alignment to an external language framework such as the CEFR might not be straightforward. As discussed earlier, by design, the description of what learners are expected to do at different performance levels of the CEFR is underspecified to allow for a wider application (Milanovic & Weir, 2010). This does not suggest that the CEFR is not a useful tool. On the contrary, it is this intentional context-free nature of the CEFR that allows for a variety of applications in designing test specifications, test tasks, and rating performance on these tasks.

The TOEFL iBT test was designed following a thorough investigation of the TLU domain, and its design included language tasks that can allow test takers to demonstrate the skills and abilities required in the TLU domain, as documented in detail in Chapelle et al. (2008). However, the connection between the content of the TOEFL iBT test and the CEFR level descriptors is evident in what both the CEFR and the TOEFL iBT test value in terms of describing English-language proficiency. In fact, the CEFR descriptors, primarily because of their context-independent nature, remain relevant to the inferences that can be made on the basis of TOEFL iBT test scores, as they are echoed in the scoring rubrics,[4] in particular the speaking rubric, given that the CEFR describes aspects of oral proficiency in more detail and on a stronger empirical basis than aspects of writing proficiency (North, 2014). As shown in the appendix, test takers at Level 3 of the speaking scoring rubric are expected to use grammar and vocabulary accurately and effectively, and mistakes do not interfere with being understood, a description closely reflecting B2 level performance on the grammatical accuracy scale of the CEFR.

To conclude, 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 that are relevant to the tasks in the TLU domain (ETS, 2014). Therefore, content alignment to the levels of an external language framework such as the CEFR can be particularly challenging, especially in the case of tests such as the TOEFL iBT test, whose design is driven by a carefully conceptualized framework of communicative language ability in

the TLU domain (Jamieson, Eignor, Grabe, & Kunnan, 2008). Nevertheless, there is a clear connection between the evidence of English proficiency valued by the TOEFL iBT test and the English skills presented in the CEFR descriptors, as reflected in the speaking scoring rubric (see the appendix). This connection suggests that although we cannot claim a one-to-one correspondence between the CEFR and the content of TOEFL iBT test, the two are sufficiently aligned to justify mapping of the TOEFL iBT test scores to the CEFR. It should also be pointed out, as we noted earlier, that standard setting remains the crucial point in the process of linking an examination to the CEFR, as it establishes a decision rule to assign students to one of the CEFR levels on the basis of their test performance (Council of Europe, 2009).

Investigating the Reasonableness of the Revised Cut Scores

To investigate whether the revised cut scores were reasonable, we conducted an investigation of university admission requirements for TOEFL iBT test scores in English speaking countries, which are popular destinations for international students. We determined the top-ranked universities based on the 2014−2015 Times Higher Education World University Rankings.[5] In support of this effort, we reviewed the admissions webpages of 155 universities in English-speaking countries starting with the top-ranked universities: 100 universities in the United States; 30 in the United Kingdom; 15 in Canada; and 10 in Australia. It is worth pointing out that at the time of our investigation some universities did not provide information about TOEFL iBT test score requirements (and could not be contacted directly due to practical constraints), for a number of possible reasons:

- Universities perceived as particularly competitive in the United Stated typically require applicants who do not speak English as their first language to take the TOEFL iBT test but without specifying a minimum score requirement.
- Minimum score requirements, in particular in universities in the United States and the United Kingdom, were found to vary by department or program of study, especially for graduate admissions; therefore, it was not possible to list a single score for undergraduate requirements for some universities and for the majority of universities for graduate admission. We decided to exclude any graduate score data from our analysis and undergraduate data from universities without a general minimum score requirement.

Table 2 presents the mean, median, minimum and maximum undergraduate score requirements identified by country for a total of 117 universities. Score requirements, in general, reflect what may be expected of students within the revised B2 range of proficiency. This finding seems to be in agreement with North's (2014, p. 62) argument that 29 descriptors at the B2 level in eight proficiency scales included in the CEFR are particularly relevant as language objectives for further and higher education.

Table 2. Summary of Undergraduate TOEFL iBT Test Score Requirements

Country	*N*	Mean TOEFL iBT total score	Median TOEFL iBT total score	Min. TOEFL iBT total score	Max. TOEFL iBT total score
Australia	7	81.14	80	79	87
Canada	14	86.36	86	79	90
UK	13	86.69	87	70	110
US	83	85.94	80	66	105

Given the results presented in Table 2, we conducted an additional analysis of the score requirements to obtain further insights into the reasonableness of the revised cut score range mapped to the B2 level of the CEFR by comparing that range to the recommended cut scores (Tannenbaum & Wylie, 2008). As shown in Table 3, the revised total score range for B2 (72−94) captures the majority of score requirements (83 out of 117 university requirements) and approximately double the number of requirements compared to the original range of scores mapped to B2 (43 out of 117 university requirements).

Table 3. Undergraduate TOEFL iBT Test Score Requirements in Relation to the Mapping of TOEFL iBT Test Scores at the B2 Level of the Common European Framework of Reference (CEFR)

Country	*N*	TOEFL iBT total score range 87−109 (original B2 cut scores)	TOEFL iBT total score range 72−94 (revised B2 cut scores)
Australia	7	1	7
Canada	14	6	14
UK	13	6	10
US	83	30	52
Total	117	43	83

Overall, the revised TOEFL iBT test score range for B2 seems to better capture the current practice in university entry requirements among the surveyed institutions. The analysis in this section provides some backing for the reasonableness of the revised CEFR cut scores, in

particular if the following is taken into consideration: With the exception perhaps of universities in the United Kingdom, requirements in the other three non-European countries are unlikely to have been directly influenced by the original CEFR mapping recommendations (Tannenbaum and Wylie, 2008) and will not be impacted by the revisions in the CEFR cut scores.

Investigating the Impact of the Revised Cut Scores

It could be argued that lowering the recommended CEFR cut scores and inevitably increasing false positive classifications could result in admission of university students who do not have sufficient English-language skills to cope with instructional demands in the English language. As discussed in Cho and Bridgeman (2012) and Bridgeman, Cho, and DiPietro (in press), English-language proficiency is a necessary but not sufficient condition for international students to succeed in universities where English is the medium of instruction. Other factors, such as subject-related knowledge and noncognitive attributes (e.g., motivation, persistence, and grit) can influence future academic performance. For this reason, English-language proficiency is typically expected to have some, but not a strong, relationship to future academic performance. With this caveat, this section investigates whether lowering the recommended cut scores could have the unintended consequence of admitting students with lower academic ability.

To investigate the impact of the revised CEFR cut scores, we utilized the data from the two previous studies that investigated the relationship between iBT scores and a grade point average (Cho & Bridgeman, 2012; Bridgeman et al., in press). Cho and Bridgeman (2012) analyzed the TOEFL iBT test scores and the cumulative grade point average (GPA) of 2,594 international students (1850 graduate and 744 undergraduate) from 10 universities in the United States. Bridgeman et al. (in press) analyzed the TOEFL iBT test scores and the overall GPA from the first three terms of 787 undergraduate international students in one university in the United States. To present the results in a more intuitive way for the purposes of our analysis, we examined the classification of test takers across two TOEFL iBT test score categories and three GPA categories. The TOEFL iBT test score categories were the revised score range mapped to the B2 level of the CEFR (72−94) and the original range of B2 scores proposed (Tannenbaum & Wylie, 2008). The GPA categories were defined based on the following rationale: below 2.6 (scores typically associated with B- or below), 2.6−3.6 (scores typically associated with B), and above 3.6 (scores typically associated with A or A-). The results are presented in Table 4.

Table 4. Classification of University Students by TOEFL iBT Test Score and Grade Point Average (GPA)

Data	Cho & Bridgeman (2012)		Bridgeman et al. (in press)		Cho & Bridgeman (2012)	
Participants	Undergraduate students		Undergraduate students		Graduate students	
TOEFL iBT total score range	87−109	72−94	87−109	72−94	87−109	72−94
N	393	332	375	398	1127	557
Below 2.6	78	75	74	92	18	11
	(19.80%)	(22.60%)	(19.70%)	(23.10%)	(1.60%)	(2.00%)
2.6−3.6	219	184	171	207	547	322
	(55.70%)	(55.40%)	(45.60%)	(52.00%)	(48.50%)	(57.80%)
Above 3.6	96	73	130	99	562	224
	(24.40%)	(22.00%)	(34.70%)	(24.90%)	(49.90%)	(40.20%)

As can be seen in Table 4, the distribution of the students based on GPA was in general comparable under the two different cut-score ranges. Some shifting of the percentages was observed among the middle (2.6−3.6) and top categories with the revised cut scores, but as was true with the originally recommended cut scores, the majority of students remained in the either the middle or top categories. Most importantly, the percentages of students in the lowest category (below 2.6) was largely unaffected by the use of the revised cut scores.

The results presented in this section should not be misunderstood as an indication that TOEFL iBT test scores do not matter in the admission process; the studies by Cho and Bridgeman (2012) and Bridgeman et al. (in press) clearly show that TOEFL iBT test scores have a meaningful relationship with academic performance as indicated by the GPA. Instead, the results suggest that using the revised score range for CEFR Level B2 will not likely result in admitting university students who do not have sufficient English-language skills to cope with instructional demands in the English language.

Discussion and Conclusions

In this research memorandum we presented the rationale behind the revision of the CEFR cut scores for the TOEFL iBT test. Moreover, we offered evidence supporting (a) the reasonableness of the revised CEFR cut scores and (b) the lack of negative impact on admissions decisions due to the classification errors associated with these revised cut scores. The supporting evidence came from external, nonassessment criteria (Kane, 2001).

One important issue we noted in this research memorandum is that content alignment to an external language framework such as the CEFR can be particularly complex, and especially so in the case of a test similar to the TOEFL iBT test that (a) does not target any one specific

level of language proficiency but rather is designed to evaluate a continuum of proficiency and (b) is modeled to reflect tasks with fidelity to the TLU domain (Chapelle et al., 2008). Nevertheless, we also demonstrated that there is construct congruence (Tannenbaum & Cho, 2014) between the TOEFL iBT test and the CEFR level descriptors. Moreover, we underlined the importance of standard setting as part of establishing the link between a language test and the CEFR, as also pointed by the Council of Europe (2009).

Another important issue in the context of mapping test scores to the CEFR levels is the interpretation of results from different assessments that claim alignment with the same CEFR level. These different assessments should not be interpreted as equivalent in terms of difficulty or content coverage, as clearly stressed in the manual (Council of Europe, 2009, p. 90). Achieving CEFR Level B1 on a general proficiency test intended for young learners and a test of English for Specific Purposes (ESP) does not mean that the scores on these two tests have the same meaning because the intended test purpose, test content, and test taking population are notably different. One way to provide more accurate information about assessment results is to provide empirically derived, test-specific performance levels and descriptors designed for a given assessment, for example by adopting a scale anchoring methodology (Garcia Gomez et al., 2007). Such levels and descriptors can be provided in addition to information about CEFR alignment (Papageorgiou, Morgan, & Becker, in press; Papageorgiou, Xi, Morgan, & So, 2015).

It also worth noting that for the purposes of simplifying the presentation of results from nonassessment criteria, we focused on the total score. However, as shown in Table 1, revised cut scores for each test section are also presented in this research memorandum. In fact, evidence is beginning to accumulate that for certain test-taker subgroups much attention should be paid to separate skill scores (Bridgeman et al., in press; Ginther, Yan, & Potts, 2015).

In conclusion, it is important to note that while it is useful to associate test scores to the CEFR levels (Kane 2012), it should not be assumed that the relationship between a language test and the CEFR levels is necessarily simple, direct, or established as a one-time event. In fact, revisions to the alignment of test scores to the CEFR levels might be required in light of additional evidence and improved understanding of the relationship among a language test, its TLU domain, and the CEFR. For this reason, various test providers in the past have revised the alignment of their scores to the CEFR levels (Cambridge Michigan Language Assessments, 2014; Pearson Education, 2012; Taylor, 2004; Taylor & Jones, 2006; University of Cambridge

ESOL Examinations, 2011). Moreover, education professionals and researchers advocate adaptation of the CEFR to meet local needs because the CEFR is a dynamic, rather than fixed, framework. Consequently, it is appropriate to periodically review and reconsider the relationship between test scores and the CEFR in light of changes in decision makers' needs and changing assessment contexts, as we demonstrate in this research memorandum because "cut scores are constructed, not found" (Zieky, 2001, p. 45). This is the case in particular for the CEFR, whose success in facilitating the meaning of test scores relies on its flexibility and the lack of any one authorized interpretation of its content (North, 2014).

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London, UK: Continuum.

Alderson, J. C. (2007). The CEFR and the need for more research. *Modern Language Journal, 91*(4), 659–663.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly, 3*(1), 3–30.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301–320.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford, UK: Oxford University Press.

Brandon, P. R. (2004). Conclusions about frequently studies modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59–88.

Bridgeman, B., Cho, Y., & DiPietro, S. (in press). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing.*

Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language education policy*. Bristol, UK: Multilingual Matters.

Cambridge Michigan Language Assessments. (2014). *MET 2009-2013 technical review.* Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/MET-TechReview-2009-2013.pdf.

Carlsen, C., & Deygers, B. (2014, April). *The B2 level and its applicability in university entrance tests.* Paper presented at the 5th ALTE International Conference, Paris, France. Retrieved from http://lirias.kuleuven.be/handle/123456789/479000

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421–442.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. London, UK: Sage.

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 Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg, France: Council of Europe.

Educational Testing Service. (2005). *Setting the final cut score.* Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/setting_final_scores.pdf

Educational Testing Service. (2014). *A guide to understanding TOEFL iBT scores.* Retrieved from https://www.ets.org/s/toefl/pdf/performance_feedback_brochure.pdf

Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem, The Netherlands: CITO.

Garcia Gomez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, *24*(3), 417–444.

Geisinger, K. F., & McCormick, C. A. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement*: *Issues and Practice*, *29*, 38–44.

Ginther, A., Yan, X., & Potts, J. (2015, March). The relationship between TOEFL and GPA: The case of Chinese students. Paper presented at the 37th Language Testing Research Colloquium, Toronto, Canada.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement*: *Issues and Practice*, *22*(1), 22–32.

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.

Hasselgreen, A. (2012). Adapting the CEFR for the classroom assessment of young learners' writing. *Canadian Modern Language Review, 69*(4), 415–435.

Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 55–96). London, UK: Routledge.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies* (pp. 106–129). Strasbourg, France: Council of Europe.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2006). Validity. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kantarcioglu, E., Thomas, C., O'Dwyer, J., & O'Sullivan, B. (2010). Benchmarking a high-stakes proficiency exam: The COPE linking project. In W. Martyniuk (Ed.), *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual* (pp. 102–116). Cambridge, UK: Cambridge University Press.

Ling, G., Wolf, M. K., Cho, Y., & Wang, Y. (2014). *English-as-a-second-language programs for matriculated students in the United States: An exploratory survey and some issues* (Research Report No. RR–14-11). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12010

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching, 39*(3), 167–190.

Martyniuk, W. (Ed.). (2010). *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual*. Cambridge, UK: Cambridge University Press.

Milanovic, M., & Weir, C. J. (2010). Series editors' note. In W. Martyniuk (Ed.), *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual* (pp. viii−xx). Cambridge, UK: Cambridge University Press.

Morgan, D. L., (2004, June). *The performance profile method (PPM): A unique standard setting method as applied to a unique population*. Paper presented at the annual meeting of the Council of Chief State School officers, Boston, MA.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.

North, B. (2014). *The CEFR in practice*. Cambridge, UK: Cambridge University Press.

North, B., Martyniuk, W, & Panthier J. (2010). Introduction: The manual for relation language examinations to the Common European Framework of Reference for Languages in the context of the Council of Europe's work on language education. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 1–17). Cambridge, UK: Cambridge University Press.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–262.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282.

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL Junior Standard scores for ESL placement decisions in secondary education. *Language Testing, 31*(2), 223–239.

Papageorgiou, S., Morgan, R., & Becker, V. (in press). Enhancing the interpretability of the overall results of an international test of English-language proficiency. *International Journal of Testing*.

Papageorgiou, S., & Tannenbaum, R. J. (in press). Situating standard setting within argument-based validity. *Language Assessment Quarterly*.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly, 12*(2), 153–177.

Pearson Education. (2012). *PTE Academic score guide.* Retrieved from http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score_Guide.pdf

Perie, M., & Thurlow, M. (2012). Setting achievement standards on assessments for students with disabilities. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 347–377). 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.

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, 233–249.

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of Angoff-based standard-setting judgments: Are item judgments and passing scores replicable across different panels of experts? *Educational Assessment, 20*(1), 66–78.

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.

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. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02120.x

Taylor, L. (2004). IELTS, Cambridge ESOL examinations and the Common European Framework. *Research Notes, 18*, 2–3. Retrieved from http://www.cambridgeenglish.org/images/23135-research-notes-18.pdf

Taylor, L., & Jones, N. (2006). Cambridge ESOL exams and the Common European Framework. *Research Notes*, *24,* 2–5. Retrieved from http://www.cambridgeenglish.org/images/22627-research-notes-24.pdf

University of Cambridge ESOL Examinations. (2011). *Using the CEFR: Principles of good practice.* Retrieved from http://www.cambridgeenglish.org/images/126011-using-cefr-principles-of-good-practice.pdf

Van Ek, J. A., & Trim, J. L. M. (1991). *Waystage 1990*. Cambridge, UK: Cambridge University Press.

Van Ek, J. A., & Trim, J. L. M. (1998). *Threshold 1990*. Cambridge, UK: Cambridge University Press.

Van Ek, J. A., & Trim, J. L. M. (2001). *Vantage*. Cambridge, UK: Cambridge University Press.

Weir, C. J. (2005). Limitations of the Common European Framework of Reference for Languages (CEFR) for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.

Wilkins, D. A. (1976). *Notional syllabuses*. Oxford, UK: Oxford University Press.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Appendix. Descriptors From Level 3 of the TOEFL iBT Speaking Scoring Rubric Compared to B2 Level Common European Framework of Reference (CEFR) Descriptors

TOEFL iBT Speaking descriptors	CEFR descriptors
• The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message (Integrated Speaking Task, Language Use, Level 3). • Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation, or pacing and may require some listener effort at times. Overall intelligibility remains good, however (Integrated Speaking Task, Delivery, Level 3).	• Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding (B2 level, Grammatical Accuracy Scale, Council of Europe, 2001, p. 114). • Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication (B2 level, Vocabulary Control Scale, Council of Europe, 2001, p. 112). • Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. (B2 level, Vocabulary Range Scale, Council of Europe, 2001, p. 112). • Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses (B2 level, Spoken Fluency Scale, Council of Europe, 2001, p. 129).

Notes

[1] https://www.gov.uk/tier-4-general-visa/knowledge-of-english

[2] We use the term *CEFR scales* (plural) to refer to what the CEFR calls "illustrative scales" (of descriptors). These scales describe various language activities and aspects of language competence across the six main levels (A1−C2) and in some cases three "plus" levels (A2+, B1+, B2+) in a table format (Council of Europe, 2001, pp. 217−225). The CEFR also contains a global scale describing overall communicative proficiency at each of the six main levels (Council of Europe, 2001, p. 24).

[3] https://www.ets.org/toefl/research/ibt_insight_series

[4] https://www.ets.org/toefl/institutions/scores/guides/

[5] http://www.timeshighereducation.co.uk/world-university-rankings/2014-15/world-ranking