INSTITUTE FOR TEST RESEARCH
AND TEST DEVELOPMENT

RESEARCH PAPERS IN ASSESSMENT

Erwin Tschirner, Olaf Bärenfänger (eds.)

Erwin Tschirner

Mapping TOEFL iBT® Scores onto the ACTFL Proficiency Guidelines

Volume 2

Bibliographische Informationen der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Die "Research Papers in Assessment" sind eine Reihe des Instituts für Testforschung und Testentwicklung e. V. (ITT), in der Forschungsergebnisse, Tagungsbeiträge und wichtige Einzeldarstellungen veröffentlicht werden.

Institut für Testforschung und Testentwicklung e.V. Leipzig
c/o Herder-Institut
Universität Leipzig
Beethovenstraße 15
04107 Leipzig
www.itt-leipzig.de

Herausgeber:
Erwin Tschirner, Universität Leipzig
Olaf Bärenfänger, Universität Leipzig

Format und Layout:
Nadja Nitsche

URN des Bandes: http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa2-734311
URN der Reihe: http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-188813

ISSN 2366-6870

Vorwort der Herausgeber

Das aussagekräftige Messen und Bewerten von Fremdsprachenkenntnissen gewinnt zunehmend an Bedeutung. Sowohl in den Bereichen Beruf und Bildung, aber auch im Privaten und nicht zuletzt im Zuge erheblicher Zu- und Abwanderungsbewegungen weltweit spielt das Beherrschen, Fördern und Evaluieren von Sprachen eine maßgebliche Rolle. Die Reihe *Research Papers in Assessment*, herausgegeben vom Vorstand des Instituts für Testforschung und Testentwicklung e. V., präsentiert aktuelle Studien zur validen und reliablen Messung von Sprachkenntnissen, zu High- und Low-Stakes-Tests, zu Testkonzepten für Unterricht und Lehrmaterialien, zu diagnostischen Testverfahren und damit verbundener individueller Sprachförderung, Sprachbedarfsanalysen und allen damit verbundenen Themen. Die Reihe erscheint als Online-Publikation, um aktuelle Forschungsergebnisse möglichst rasch interessierten WissenschaftlerInnen, Lehrkräften und mit dem Testen von Fremdsprachenkenntnissen betrauten Institutionen zugänglich zu machen und diese in die Testpraxis umsetzen zu können.

Die Herausgeber

Erwin Tschirner

Olaf Bärenfänger

Mapping TOEFL iBT® Scores onto the ACTFL Proficiency Guidelines

Erwin Tschirner

Contents

Acknowledgments

Acknowledgments are gratefully made to ACTFL for initiating and supporting the study; LTI for generously supporting the research financially and administratively; the instructors and students at all participating universities who generously gave their time; and to Margaret E. Malone from the ACTFL Center for Assessment Research and Development (CARD) for many insightful comments and suggestions on this technical report.

Introduction

The purpose of this study was to establish a crosswalk between the Test of English as a Foreign Language (TOEFL iBT®) and four ACTFL Assessments to help examinees and institutions of higher education (IHE) to better understand the correspondences between TOEFL iBT scores and examinees' functional proficiency, i.e., their ability to use functional English in real-world academic and social situations. The ACTFL Proficiency Guidelines 2012 describe what an individual can do consistently with his or her language abilities while listening and reading and in speaking and writing. Because the ACTFL Proficiency Guidelines 2012 provide a developmental perspective, i.e., what an examinee is able to do now and will be able to do at the next higher level, test results may also be used to determine linguistic areas to be targeted to improve students' proficiency.

The results of this study may also benefit IHEs by providing a research-based interpretation of how TOEFL iBT scores relate to functional language ability in an English-language context. In addition, IHEs receive more fine-grained information about their students' abilities, including diagnostic feedback to pass on to their students. Moreover, IHEs will learn how ACTFL Assessments might be used to further their own mission with respect to admission and placement and ultimately to professional career goals. Furthermore, they may be able to reflect on and (re-)evaluate their existing minimal TOEFL iBT scores for admission purposes.

Review of the Literature

There are several major frameworks or guidelines that provide guidance to how world languages are learned, taught, and assessed for functional purposes. Two of the most widely known are the ACTFL Proficiency Guidelines based on the US government Interagency Language Roundtable (ILR) proficiency level descriptors and the Common European Framework of Reference for Languages (CEFR). ACTFL and the Council of Europe (CoE) have collaborated on establishing correspondences between these two systems. In addition, international test publishers such as ETS have found the need to "map" their tests to them. This section will summarize the existing research on the mapping of TOEFL iBT test scores and ACTFL proficiency levels to the CEFR.

Tannenbaum and Wylie (2008) mapped TOEFL iBT scores onto the CEFR following standard-setting methods using expert judgment. A modified Angoff approach was used for the selected-response reading and listening sections (cf. Impara & Plake, 1997) and a modified examinee selection approach was employed for the constructed-response writing and speaking sections (cf. Hambleton, Jaeger, Plake, & Mills, 2000). Based on these standard-setting methods, links were established between the scores of each subtest of the TOEFL iBT and CEFR levels. These links were represented in the form of cut scores. A cut score is the minimum score that experts judge as necessary for a given level.

For the TOEFL iBT, the experts found that the listening and reading sections of the TOEFL iBT were too demanding for test-takers at the A1 and A2 levels. The writing section was considered too difficult for candidates at the A1 level. In addition, the judges were of the opinion that the listening, speaking, and writing sections were not challenging enough at the C2 level. Accordingly, cut scores were established from B1 to C2 for reading, from B1 to C1 for listening, from A1 to C1 for speaking, and from A2 to C1 for writing (see Table 1).

Table 1: TOEFL iBT® Cut Scores for CEFR Levels (2008)

	Reading (0-30)	**Listening (0-30)**	**Speaking (0-30)**	**Writing (0-30)**
C2	29			
C1	28	26	28	28
B2	22	21	23	21
B1	8	13	19	17
A2			13	11
A1			8	

Table 1 shows a one-point difference between C1 and C2 for reading and a large gap between B1 and B2 for both reading and listening. Speaking and writing scores seem to be distributed more evenly. Generally, TOEFL iBT scores in the low twenties seem to be associated with the B2 level, whereas scores in the high twenties correspond to the C1 level.

In 2014, ETS revisited the 2008 cut score recommendations above because of feedback by users and decision makers,

mostly universities in the U.K. and other European countries, which use CEFR levels for admission decisions (cf. Papageorgiou, Tannenbaum, Bridgeman, & Cho, 2015). Many of these decision makers felt that the cut scores were too high, particularly for the B2 level, which appears to be the most common requirement for admission into European universities (cf. Carlsen & Deygers, 2014). Consequently, ETS lowered the cut scores established in 2008 by two standard errors of measurement (SEM) to reduce the likelihood of making false-negative admission decisions because they claimed that many institutions are more in favor of giving examinees the benefit of the doubt rather than making sure that everybody was functioning at the level required. Table 2 shows the revised recommended cut scores.

Table 2: TOEFL iBT® Cut Scores for CEFR Levels (2015)

	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		

While the new cut scores may be reasonable for the B2 level, and possibly, the C1 level, assuming a TOEFL iBT score of 25 to correspond to the C2 level does not appear to be justified because it retains the one-point difference between C1 and C2

seemingly ignoring the vastly expanded proficiency of C2 over C1 readers. Moreover, associating scores of 4 and 9 with B1 in reading and listening, respectively, appears to underestimate the proficiency of B1 readers and listeners. In the absence of additional empirical evidence, this unilateral lowering of the cut scores by two SEMs, therefore, may not be justified.

Additional evidence that the speaking score may be inflated comes from two studies looking at the speaking section of the TOEFL iBT and the Test of Spoken English (TSE). Wylie and Tannenbaum (2006) established cut scores for international teaching assistants for speaking. Using standard-setting methods, they set the cut score for minimally acceptable speaking skills, i.e., the proficiency required for the lowest level of contact with undergraduate students, at 23, while they put the cut score that corresponded to a score of 50 on the TSE to 26. A score of 50 on the TSE was considered a robust level of speaking proficiency often required for graduate student admission. In another study, Wylie and Tannenbaum (2005) associated a TSE score of 45 with the B1 level and 55 with the C1 level. A score of 50 on the TSE, i.e., a score of 26 on the speaking section of the TOEFL iBT, would, therefore, fall somewhere between B1 and C1, possibly B2, and would not be associated with C1 as Papageorgiou et al. (2015) state.

Bärenfänger and Tschirner (2012) established correspondences between ACTFL speaking proficiency levels and the CEFR. They linked the ACTFL Oral Proficiency Interview by computer (OPIc) to the CEFR, following the benchmarking procedure established by the Council of Europe

(cf. Council of Europe, 2009) to link assessments to the CEFR. The benchmarking was conducted with six expert raters of CEFR oral proficiency tests in German. They were asked to assign CEFR ratings to a total of 54 German OPIc and OPI samples with official ACTFL ratings. Interrater reliability was very high with Kendall's concordance coefficient $W = 0.96$ ($p < .001$). Correlation and agreement measures between ACTFL and CEFR ratings were also very high: Spearman's *rho* = 0.966 and Goodman Kruskal's *gamma* = 0.968 (both at $p < .01$). Table 3 shows the correspondences between ACTFL and CEFR ratings.

Table 3: Correspondences Between ACTFL and CEFR Ratings of OPIc and OPI Samples

ACTFL	NH	IL	IM	IH	AL	AM	AH	S
CEFR	A1	A2	B1	B1+	B2	B2+	C1	C2

To be able to make finer distinctions, the CEFR uses plus sublevels such as B1+ etc. (cf. Council of Europe, 2001). The use of 1 and 2, e.g., A1.1 and A1.2 etc. (see below), is another convention to distinguish between base and plus levels.

Based on Bärenfänger and Tschirner (2012) and other studies, ACTFL (2016) published official correspondences between ACTFL and CEFR ratings and ACTFL assessments. Table 4 shows these correspondences for all four skills.

Table 4: Correspondences Between CEFR and ACTFL Levels

CEFR	ACTFL Reading and Listening	ACTFL Speaking and Writing
C2	Distinguished	Superior
C1.2	Superior	Advanced High
C1.1	Advanced High	Advanced High
B2.2	Advanced Mid	Advanced Mid
B2.1	Advanced Mid	Advanced Low
B1.2	Advanced Low	Intermediate High
B1.1	Intermediate High	Intermediate Mid
A2	Intermediate Mid	Intermediate Low
A1.2	Intermediate Low	Novice High
A1.1	Novice High	Novice High

Note that Table 4 shows slightly different correspondences for the receptive and the productive skills. In the next sections, the methods and results of the present study will be presented.

Methods

Participants

A total of 234 examinees participated in the study. They were students at the following universities: Cornell University, Georgetown University, Miami Dade College, Michigan State University, State University of New York at Plattsburgh, Teachers' College of New Jersey, University of Hartford, University of Utah, and Yale University.

53.8 % of the examinees were female, while 46.2 % were male. 56.8 % of the examinees were graduate students, 32.2 % were undergraduate students, and 10.7 % were exchange students or students enrolled in non-degree programs such as teacher education. 37.2 % of the examinees had Chinese, 16.7 % Portuguese, 15.4 % Arabic, 7.7 % Spanish, 3.4 % Korean, and 3.0 % Thai as their first language. Other first languages were Bengali, English, French, German, Gujarati, Hebrew, Hindi, Italian, Japanese, Kirundi, Malay, Malayalam, Norwegian, Punjabi, Russian, Turkish, Urdu, and Vietnamese. The average number of years examinees had studied and/or used English was 11.55 years (SD = 6.29, Min = 1, Max = 37). Note that the majority of the students were graduate students (57 %) and most of them had Chinese as their first language (37 %).

Instruments

The ACTFL assessments consisted of the ACTFL L&Rcat, a machine-scored computer-adaptive test of listening and reading proficiency; the ACTFL OPIc, an online speaking test with prerecorded oral prompts, which is blindly double-rated by human raters; and the ACTFL WPT, an online writing test with written prompts, which is also blindly double-rated by human raters.

The ACTFL L&Rcat is a computer-adaptive test designed to measure the listening and reading proficiency of examinees in English. It currently has an item bank consisting of 1,500 items. All items were calibrated in 20 separate pilot studies with an overall total of more than 4,000 examinees to determine difficulty values measured in logits for each individual item. The L&Rcat algorithm selects appropriate items for examinees on the basis of the correctness of their previous responses and calculates a final person ability value also measured in logits at the end of the test. Person ability values are subsequently rendered as ACTFL sublevels.

The Test of English as a Foreign Language (TOEFL iBT) has four sections: reading, listening, speaking, and writing. The reading section consists of 36–56 questions and the listening section consists of 34–51 questions. Both sections are scored by computer. The speaking section has 6 tasks, which are scored by human raters. The writing section has 2 tasks, which are scored either by human raters or by a combination of human raters scoring content and meaning and automated scoring for

linguistic features. For each section, raw scores are converted to scaled scores of 0–30.

Examinees also completed a background survey to provide biographical information and information on their English language background. The participating universities provided student TOEFL iBT scores and the date the TOEFL iBT was taken.

Data Collection

Data collection took place between July 2015 and July 2018. A total of 202 ACTFL reading, 203 ACTFL listening, 58 ACTFL writing, and 56 ACTFL speaking assessments were administered to foreign students with known TOEFL iBT scores admitted to U.S. colleges and universities.

Results

Reading Proficiency

A total of 202 ACTFL reading assessments were administered. The reading assessment took, on average, 33:50 minutes (*SD* = 9:12; Min = 2:04; Max = 51:59). Three results were removed because students speeded through the test and could not possibly have read all of the texts (test duration less than 13 minutes). Another four results were removed on account of being outliers. All three outliers identified as such by SPSS while using ACTFL sublevel as the category axis and TOEFL iBT score as the variable in a box plot were removed in addition to one extreme outlier identified by SPSS in the box plot with TOEFL iBT score as the category axis and ACTFL sublevel as variable. The following analysis, accordingly, was based on 195 ACTFL reading assessments.

The ACTFL proficiency levels assessed ranged from Novice Low (NL) to Superior (S). Figure 1 and Table 5 present the distribution of ACTFL reading proficiency levels of 195 participants.

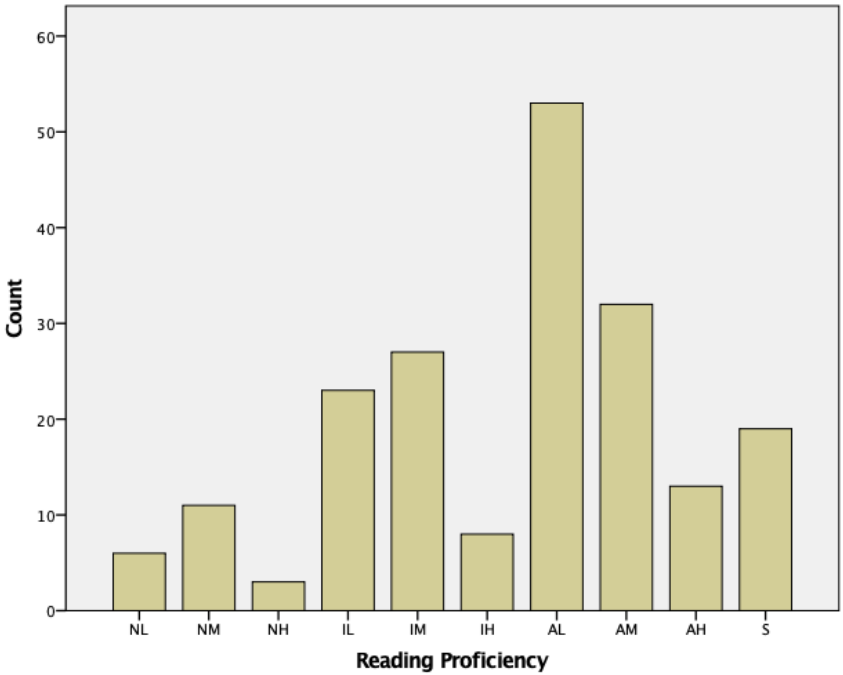


Figure 1: Distribution of ACTFL Reading Proficiency Levels
Note: NL = Novice Low; NM = Novice Mid; NH = Novice High; IL = Intermediate Low; IM = Intermediate Mid; IH = Intermediate High; AL = Advanced Low; AM = Advanced Mid; AH = Advanced High; S = Superior

Figure 1 shows that the results peak at Advanced Low (AL) and slope downward towards Intermediate and Novice on the left and Superior on the right, indicating a relatively normal distribution.

Table 5: Distribution of ACTFL Reading Proficiency Levels

ACTFL Numeric	ACTFL Level	Frequency	Percent	Cumulative Percent
1	NL	6	3.1	3.1
2	NM	11	5.6	8.7
3	NH	3	1.5	10.3
4	IL	23	11.8	22.1
5	IM	27	13.8	35.9
6	IH	8	4.1	40.0
7	AL	53	27.2	67.2
8	AM	32	16.4	83.6
9	AH	13	6.7	90.3
10	S	19	9.7	100.0
Total		195	100	

Table 5 shows that 10.3 % of the participants were Novice in reading, almost 30 % were Intermediate, 50.3 % were Advanced, and almost 10 % were Superior. Figure 2 and Table 6 show the distribution of the TOEFL iBT reading scores for the 195 participants who took the ACTFL reading proficiency assessment.

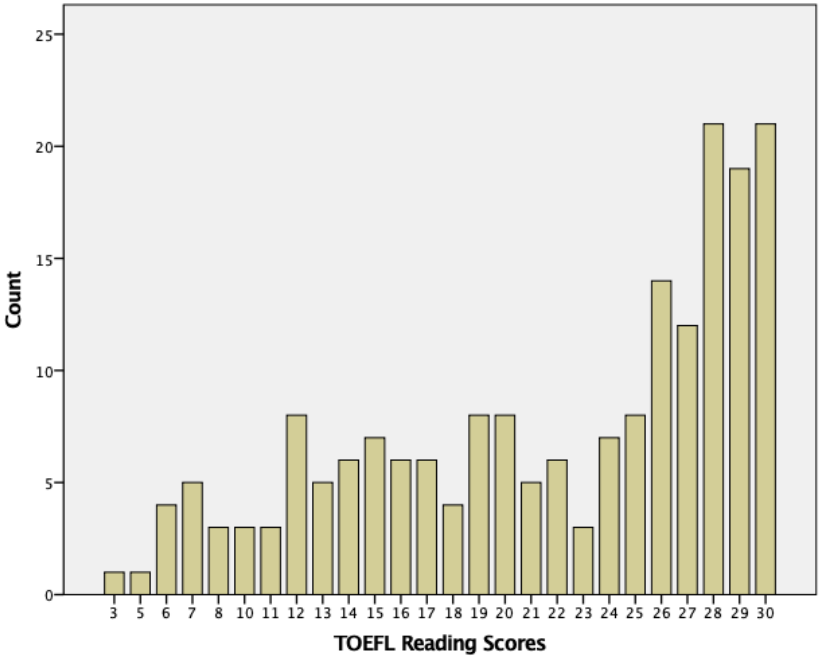


Figure 2: Distribution of TOEFL iBT® Reading Scores

Figure 2 shows that the TOEFL iBT scores peak at the three highest scores and slope down to the left, indicating a ceiling effect.

Table 6: Distribution of TOEFL iBT® Reading Scores

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
3	1	0.5	0.5
5	1	0.5	1
6	4	2.1	3.1
7	5	2.6	5.6
8	3	1.5	7.2
9	1	0.5	7.7
10	3	1.5	9.2
11	3	1.5	10.8
12	8	4.1	14.9
13	5	2.6	17.4
14	6	3.1	20.5
15	7	3.6	24.1
16	6	3.1	27.2
17	6	3.1	30.3
18	4	2.1	32.3
19	8	4.1	36.4
20	8	4.1	40.5
21	5	2.6	43.1
22	6	3.1	46.2
23	3	1.5	47.7
24	7	3.6	51.3
25	8	4.1	55.4
26	14	7.2	62.6
27	12	6.2	68.7

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
28	21	10.8	79.5
29	19	9.7	89.2
30	21	10.8	100
Total	195	100	

TOEFL iBT reading scores of 22–30 are considered to be high, 15–21 intermediate, and 0–14 low (cf. Educational Testing Service, 2014). Table 6 shows that more than 56.9 % of the results consisted of high scores. In fact, 52.3 % of the results were in the top 20 % of scores (24 points or more) and 37.4 % of the results were in the top 10 % (27 points or more). Both Figure 2 and Table 6 indicate a ceiling effect. Table 7 provides the descriptive statistics of the ACTFL and TOEFL iBT reading results.

Table 7: Descriptive Statistics of ACTFL Reading Proficiency Levels and TOEFL iBT® Reading Scores

	ACTFL Levels	TOEFL iBT Score
Possible Range	1–10 (NL to S)	1–30
Observed Range	1–10 (NL to S)	3–30
Median	7	24
Mean	6.39	21.67
Standard Error of the Mean	0.17	0.53
Standard Deviation	2.32	7.34

To align ACTFL ratings and TOEFL iBT scores, logits were used. The L&Rcat measures person ability in logits on the basis of

item difficulties also measured in logits, using item response theory (IRT). Logits, ranging from -4 to 4, provide a more fine-grained measure than ACTFL sublevels. The correlation between reading ability logits and TOEFL iBT scores was high: Pearson's $r = 0.796$, 2-tailed, $p > 0.01$, $N = 195$. Figure 3 plots TOEFL iBT reading scores and ACTFL person ability logits as determined by the reading section of the ACTFL L&Rcat.

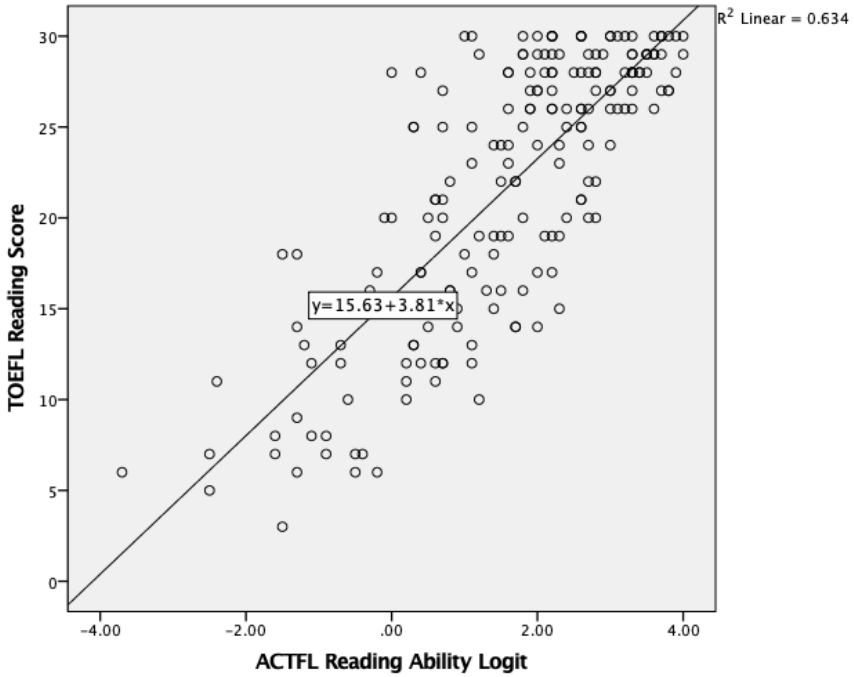


Figure 3: Scatter Plot of ACTFL Person Reading Ability Logits and TOEFL iBT® Reading Scores

Figure 3 shows the relationship between ACTFL reading ability logits and TOEFL iBT scores. ACTFL reading ability logits accounted for 63.4 % of the variance of the TOEFL iBT reading score (R^2 = 0.634). This is a very large effect. (Effect sizes above R^2 = 0.25 are considered large.) Figure 4 shows a P–P plot of the standardized residuals examining the assumption of normal distribution of the ACTFL and TOEFL iBT reading data.

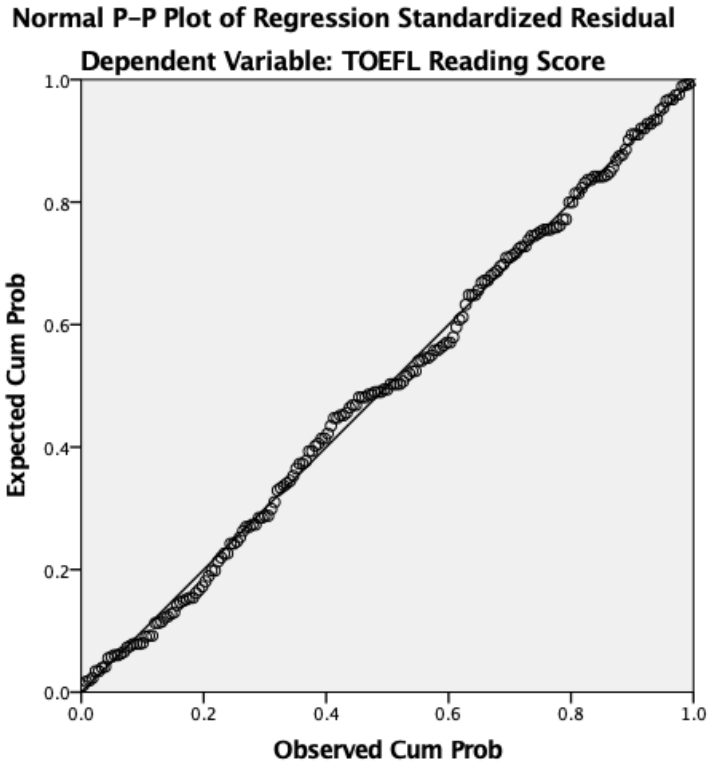


Figure 4: P–P Plot of the Standardized Residuals of ACTFL Person Reading Ability Logits and TOEFL iBT® Reading Scores

The P–P plot by and large shows a linear relationship between TOEFL iBT reading scores and ACTFL reading proficiency logits. Equipercentile scale mapping was used to establish correspondences between TOEFL iBT scores and ACTFL ratings. TOEFL iBT scores and ACTFL logits were aligned using cut points for 99 equal groups. Table 8 shows the relationship between TOEFL iBT reading scores and ACTFL reading logits and reading proficiency levels. ACTFL reading proficiency levels were converted from their corresponding person ability logits. Note that ACTFL reading proficiency levels correspond to a range of logit values and, consequently, to a range of TOEFL iBT scores.

Table 8: TOEFL iBT® Reading Scores and ACTFL Reading Ability Logits and Proficiency Ratings

TOEFL iBT Score	Logits	ACTFL	TOEFL iBT Score	Logits	ACTFL
3	−2.5242	NL	17	0.7000	IM
4	−2.5000	NL	18	0.9700	IM
5	−2.4985	NL	19	1.1000	IM
6	−1.5500	NL	20	1.4400	IH
7	−1.2200	NM	21	1.5300	AL
8	−1.1000	NM	22	1.6000	AL
9	−0.9000	NM	23	1.7000	AL
10	−0.7000	NH	24	1.8000	AL
11	−0.6400	NH	25	1.9000	AL
12	−0.3500	IL	26	2.0800	AL
13	0.0700	IL	27	2.3000	AL

TOEFL iBT Score	Logits	ACTFL	TOEFL iBT Score	Logits	ACTFL
14	0.3000	IL	28	2.6000	AM
15	0.4100	IL	29	3.0000	AM
16	0.6000	IM	30	3.4400	AH

Table 8 shows that an ACTFL rating of NL corresponds to TOEFL iBT scores 3–6; ACTFL NM to TOEFL iBT scores 7–9; ACTFL NH to TOEFL iBT scores 10–11; ACTFL IL to TOEFL iBT scores 12–15; ACTFL IM to TOEFL iBT scores 16–19; ACTFL IH to a TOEFL iBT score of 20; ACTFL AL to TOEFL iBT scores 21–27; ACTFL AM to TOEFL iBT scores 28–29; and ACTFL AH to a TOEFL iBT score of 30.

Because equipercentile scale mapping is a robust method for establishing correspondences, the lowest score of a particular range was generally used as the suggested cut score. In a few instances, the cut score was modified because of the score interpretations used by ETS (high, intermediate, and low) and the results of their standard setting studies (see below). Table 9 shows the suggested correspondences between ACTFL reading proficiency levels and TOEFL iBT reading scores based on the present study. Because levels below ACTFL Intermediate are unlikely to be of interest to college admissions decision makers, the ACTFL Novice levels are excluded.

Table 9: Correspondences between ACTFL Reading Proficiency Levels and TOEFL iBT® Reading Scores

ACTFL	IL	IM	IH	AL	AM	AH
TOEFL iBT	12	15	20	22	28	30

For IL, the lowest score of the IL range of 12–15 was used. IL is associated with CEFR A2 (ACTFL, 2016). Both the original and revised ETS crosswalks (cf. Tannenbaum & Wylie, 2008; Papageorgiou et al., 2015) associate the A2 level with TOEFL iBT scores below 12. The lowest score of the IM range was 16. IM corresponds to CEFR B1, which according to the revised ETS crosswalk is associated with TOEFL iBT scores of 4–17. The original crosswalk associated B1 with TOEFL iBT scores of 8–21. Moreover, ETS considers TOEFL iBT scores of 15–21 as intermediate. TOEFL iBT reading scores of 22–30 are considered high, 15–21 intermediate, and 0–14 low (cf. Educational Testing Service, 2014). The cut score for IM, therefore, was set to 15, the lowest TOEFL iBT intermediate score. For IH, the only TOEFL iBT score was 20, which was selected as the cut score.

For AL, the lowest score was 22. AL corresponds to CEFR B2, which has a TOEFL iBT range of 22–27 in the original crosswalk and a TOEFL iBT range of 18–23 according to the revised ETS crosswalk. Moreover, ETS considers TOEFL iBT scores of 22 and higher as high scores. The cut score for AL, therefore, was set to 22, the lowest TOEFL iBT high score. For AM, the lowest score of 28 was selected as the cut score. The ACTFL CEFR crosswalk associates AM with the upper half of B2. This is supported by the original crosswalk, which considered 27 as the highest B2 score, just shy of 28. For AH, the only TOEFL iBT

score was 30, which was selected as the cut score. AH corresponds to CEFR C1, which was associated with TOEFL iBT scores of 28–30 in the original TOEFL iBT CEFR crosswalk.

Listening Proficiency

A total of 203 listening assessments were administered. The listening assessment took, on average, 30:45 minutes (*SD* = 5:51; Min = 17:55; Max =44:20). As all examinees had to listen to the passages before they could select their responses, no results were removed on account of speeding through the test. In addition, no outliers needed to be removed. Accordingly, all 203 results were used for the present analysis. Figure 5 and Table 10 present the distribution of ACTFL listening proficiency levels for 203 participants.

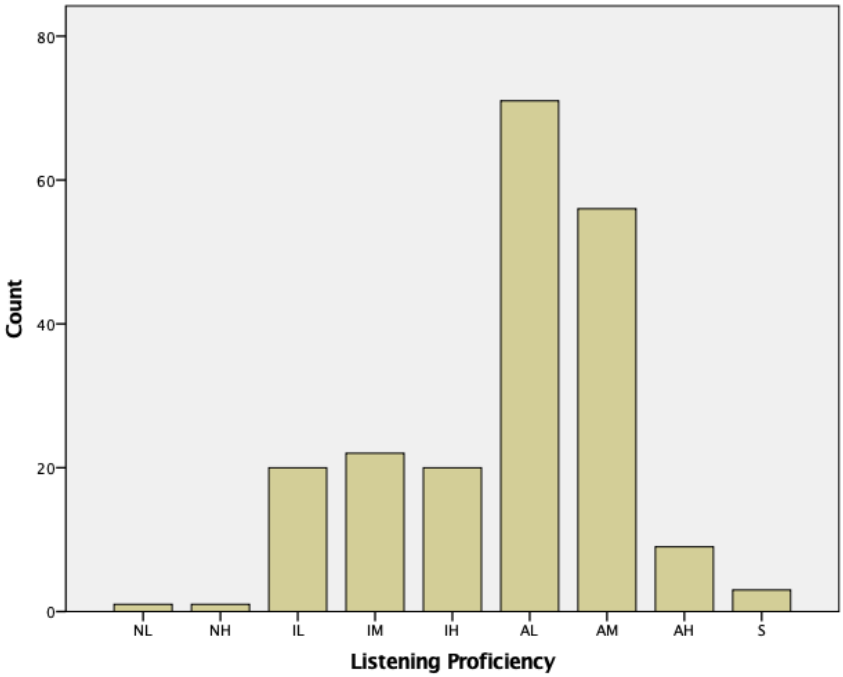


Figure 5: Distribution of ACTFL Listening Proficiency Levels
Note: NL = Novice Low; NH = Novice High; IL = Intermediate Low; IM = Intermediate Mid; IH = Intermediate High; AL = Advanced Low; AM = Advanced Mid; AH = Advanced High; S = Superior

Figure 5 shows that the assessment results peak at AL and slope downward towards Intermediate and Novice on the left and Superior on the right, indicating a relatively normal distribution.

Table 10: Distribution of ACTFL Listening Proficiency Levels

ACTFL Numeric	ACTFL Level	Frequency	Percent	Cumulative Percent
1	NL	1	0.5	0.5
3	NH	1	0.5	1.0
4	IL	20	9.9	10.8
5	IM	22	10.8	21.7
6	IH	20	9.9	31.5
7	AL	71	35.0	66.5
8	AM	56	27.6	94.1
9	AH	9	4.4	98.5
10	S	3	1.5	100.0
Total		203	100.0	

Table 10 shows that only 1 % of the participants were Novice in listening, approximately 30 % were Intermediate, 67 % were Advanced, and only 1.5 % were Superior. Figure 6 and Table 11 show the distribution of the TOEFL iBT listening scores for the 203 participants who took the ACTFL listening proficiency assessment.

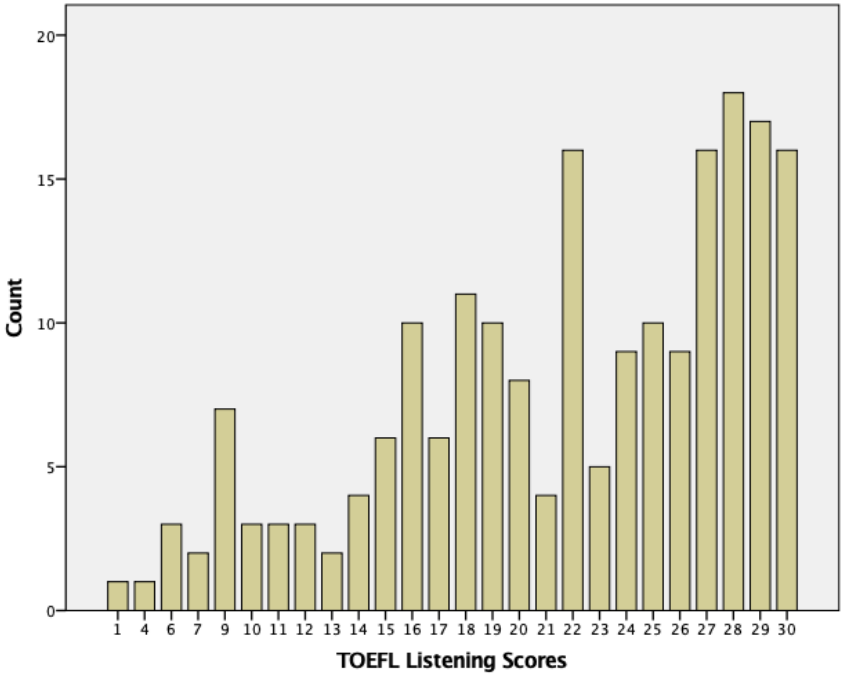


Figure 6: Distribution of TOEFL iBT® Listening Scores

Figure 6 shows that the TOEFL iBT scores peak at the highest scores (27–30) and slope down to the left, indicating a ceiling effect.

Table 11: Distribution of TOEFL iBT® Listening Scores

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
1	1	0.5	0.5
4	3	1.5	2.0

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
6	3	1.5	3.4
7	2	1.0	4.4
9	7	3.4	7.9
10	4	2.0	9.9
11	3	1.5	11.3
12	3	1.5	12.8
13	2	1.0	13.8
14	4	2.0	15.8
15	6	3.0	18.7
16	10	4.9	23.6
17	6	3.0	26.6
18	11	5.4	32.0
19	10	4.9	36.9
20	8	3.9	40.9
21	4	2.0	42.9
22	16	7.9	50.7
23	5	2.5	53.2
24	9	4.4	57.6
25	10	4.9	62.6
26	9	4.4	67.0
27	16	7.9	74.9
28	18	8.9	83.7
29	17	8.4	92.1
30	16	7.9	100.0
Total	203	100.0	

TOEFL iBT listening scores of 22–30 are considered high, 14–21 intermediate, and 0–13 low scores (cf. Educational Testing Service, 2014). Table 11 shows that 57.1 % of the participants had high scores. In fact, 33 % of the results were in the top 10 % of scores (27–30 points). Both Figure 6 and Table 11 point to a ceiling effect. Table 12 provides the descriptive statistics of the ACTFL and TOEFL iBT listening results.

Table 12: Descriptive Statistics of ACTFL Listening Proficiency Levels and TOEFL iBT® Listening Scores

	ACTFL Levels	TOEFL iBT Score
Possible Range	1–10 (NL to S)	1–30
Observed Range	1–10 (NL to S)	1–30
Median	7	22
Mean	6.75	21.47
Standard Error of the Mean	0.10	0.49
Standard Deviation	1.49	6.93

To align ACTFL ratings and TOEFL iBT scores, person ability listening logits were used. The correlation between ACTFL ratings and TOEFL iBT listening scores was high: Pearson's $r = 0.708$, 2-tailed, $p < .01$, $N = 203$. Figure 7 plots TOEFL iBT listening scores and ACTFL person listening ability logits as determined by the listening section of the ACTFL L&Rcat.

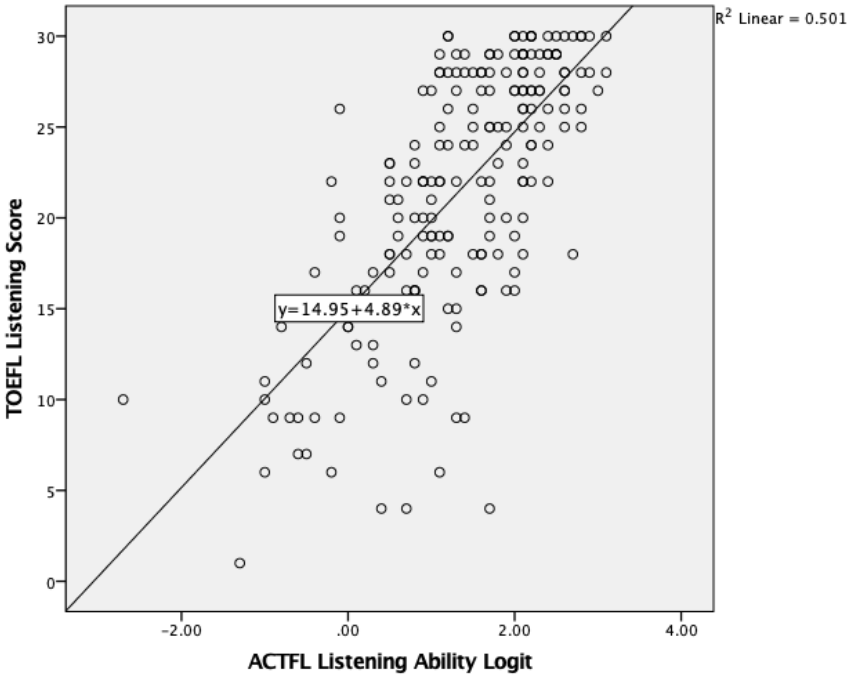


Figure 7: Scatter Plot of TOEFL iBT® Listening Scores and ACTFL Listening Proficiency Logits

Figure 7 shows the relationship between ACTFL listening ability logits and TOEFL iBT listening scores. ACTFL proficiency logits accounted for 50.1 % of the variance of the TOEFL iBT listening score ($R^2 = 0.501$). This is a very large effect. Effect sizes above $R^2 = 0.25$ are considered to be large. Figure 8 shows a P–P plot of the standardized residuals examining the assumption of normal distribution of the ACTFL and TOEFL iBT listening data.

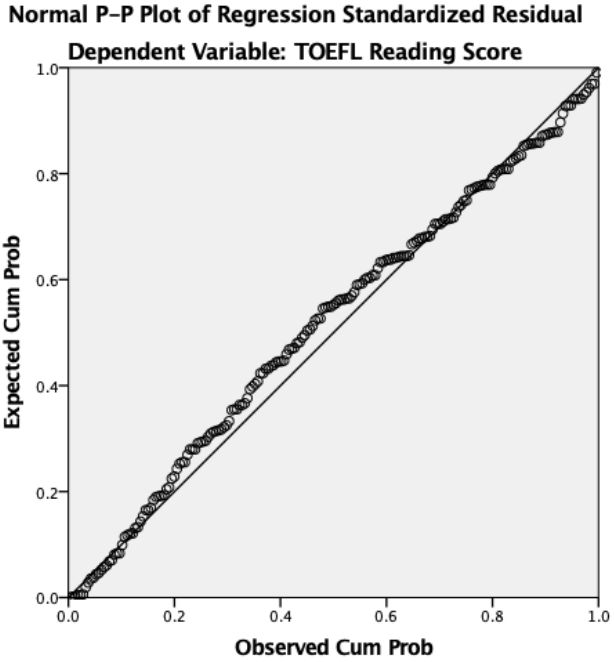


Figure 8: P–P Plot of the Standardized Residuals of ACTFL Listening Proficiency Logits and TOEFL iBT® Listening Scores

The P–P plot by and large shows a linear relationship between TOEFL iBT listening scores and ACTFL listening proficiency logits. Given the linear relationship, and both high correlations and effect sizes between the two variables, equipercentile scale alignment was used to establish correspondences between TOEFL iBT scores and ACTFL logits and ratings. TOEFL iBT scores and ACTFL logits were aligned using cut points for 99 equal groups. Table 13 shows the relationship between TOEFL iBT listening scores and ACTFL listening proficiency levels.

ACTFL listening proficiency levels were converted from their corresponding person ability logit. Note that ACTFL listening proficiency levels correspond to a range of logit values and, consequently, to a range of TOEFL iBT scores.

Table 13: TOEFL iBT® Listening Scores, ACTFL Listening Proficiency Logits and ACTFL Ratings

TOEFL iBT	Logits	ACTFL	TOEFL iBT	Logits	ACTFL
4	-1.2909	NH	17	0.7000	IH
5	-1.1463	NH	18	0.8000	IH
6	-0.9939	NH	19	1.0000	AL
7	-0.7909	IL	20	1.1000	AL
8	-0.6463	IL	21	1.2000	AL
9	-0.6000	IL	22	1.2000	AL
10	-0.2925	IL	23	1.5000	AL
11	-0.1000	IM	24	1.6000	AL
12	0.0000	IM	25	1.7000	AL
13	0.0364	IM	26	1.8000	AL
14	0.1394	IM	27	2.0000	AM
15	0.3150	IM	28	2.1000	AM
16	0.5000	IM	29	2.3000	AM
			30	2.6000	AM

Table 13 shows that an ACTFL rating of NH corresponds to TOEFL iBT scores 4-6; ACTFL IL to TOEFL iBT scores 7-10; ACTFL IM to TOEFL iBT scores 11-16; ACTFL IH to TOEFL iBT scores 17-18; ACTFL AL to TOEFL iBT scores 19-26; and ACTFL AM to TOEFL iBT scores of 27-30.

Because equipercentile scale mapping is a robust method for establishing correspondences, the lowest score of a particular range was generally used as the suggested cut score. In a few instances, the cut score was modified because of the score interpretations used by ETS (high, intermediate, and low) and the results of their standard setting studies. Table 14 shows the suggested correspondences between ACTFL listening proficiency levels and TOEFL iBT listening scores based on the present study. Because levels below ACTFL Intermediate are unlikely to be of interest to college admissions decision makers, the ACTFL Novice levels are excluded.

Table 14: Correspondences between ACTFL Listening Proficiency Levels and TOEFL iBT® Listening Scores

ACTFL	IL	IM	IH	AL	AM	AM
TOEFL iBT	7	14	17	22	27	30

The lowest score of the IL range of 7–10 was used as the cut score. IL is associated with CEFR A2 (cf. ACTFL, 2016). The original ETS CEFR crosswalk associates TOEFL iBT scores below 13 with the A2 level, while the revised crosswalk associates the A2 level with TOEFL iBT scores below 8. For IM, the median was used. The median of the IM range of 11–16 was 13.5, rounded to 14. IM corresponds to CEFR B1, which according to the revised ETS crosswalk is associated with TOEFL iBT scores of 9–16. The original crosswalk associated B1 with TOEFL iBT scores of 13–20. Moreover, ETS considers TOEFL iBT scores of 14–21 as intermediate. TOEFL iBT listening scores of 22–30 are considered high, 14–21 intermediate, and 0–13 low scores (cf.

Educational Testing Service, 2014). The cut score for IM, therefore, was set to 14, the lowest TOEFL iBT intermediate score. IH was associated with TOEFL iBT scores of 17 and 18. IH corresponds to the upper half of B1. The original ETS crosswalk associated scores of 13–20 with B1, while the revised crosswalk associated a score of 17 with B2. The lower of the two TOEFL iBT scores, i.e., 17, therefore, was selected as the cut score.

ACTFL AL was associated with TOEFL iBT scores of 19–26. AL corresponds to CEFR B2, which had a TOEFL iBT range of 21–25 in the original crosswalk and a TOEFL iBT range of 17–21 according to the revised ETS crosswalk. In addition, ETS considers TOEFL iBT scores of 22 and higher as high scores. While the rounded median for AL was 23, a score of 22 was selected as the cut score, just shy of the B2 range according to the revised TOEFL iBT CEFR crosswalk. For AM, the lowest score of 27 was selected as the cut score. The ACTFL CEFR crosswalk associates AM with the upper half of B2. This is supported by the original crosswalk, which considered 25 as the highest B2 score, two points shy of 27. A TOEFL iBT score of 30 was still associated with AM. Therefore, the score range for AM includes TOEFL iBT scores from 27 to 30.

Speaking Proficiency

A total of 55 participants had both TOEFL iBT speaking scores and ACTFL speaking ratings (OPIc). Figure 9 and Table 15 present the distribution of ACTFL speaking proficiency levels

for the 55 participants who took the ACTFL Oral Proficiency Interview by computer (OPIc).

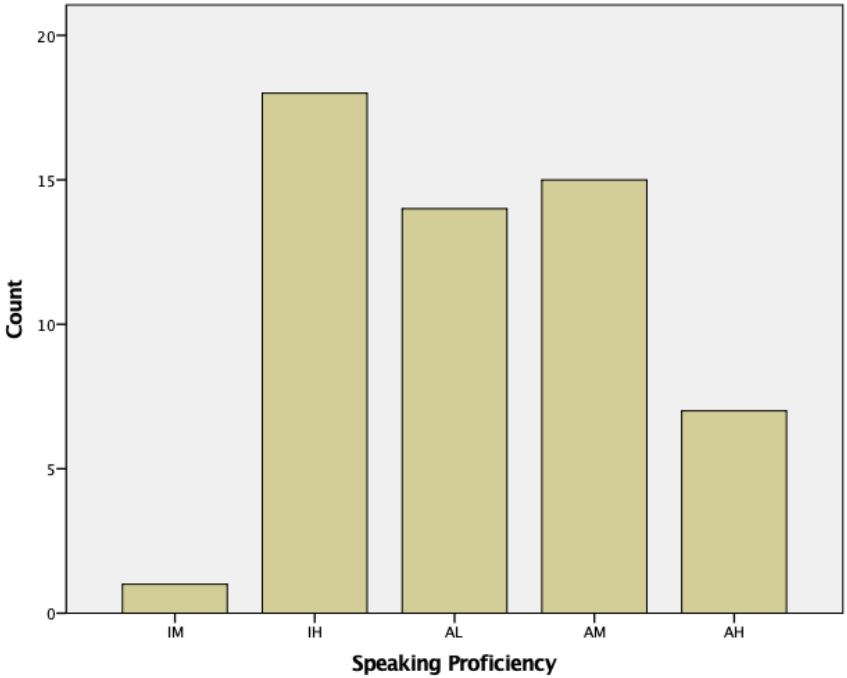


Figure 9: Distribution of ACTFL Speaking Proficiency Levels
Note: IM = Intermediate Mid; IH = Intermediate High;
AL = Advanced Low; AM = Advanced Mid; AH = Advanced High

Figure 9 shows that the assessment results peak at Intermediate High (IH) and that they slope downwards to Advanced Low (AL), Mid (AM), and High (AH) on the right, indicating a right-skewed distribution.

Table 15: Distribution of ACTFL Speaking Proficiency Levels

ACTFL Numeric	ACTFL Level	Frequency	Percent	Cumulative Percent
5	IM	1	1.8	1.8
6	IH	18	32.7	34.5
7	AL	14	25.5	60.0
8	AM	15	27.3	87.3
9	AH	7	12.7	100.0
Total		55	100	

Table 15 shows that the largest number of participants were Intermediate High (IH); approximately one third were Intermediate and two thirds were Advanced. Figure 10 and Table 16 show the distribution of the TOEFL iBT speaking scores for the 55 participants who took the ACTFL OPIc.

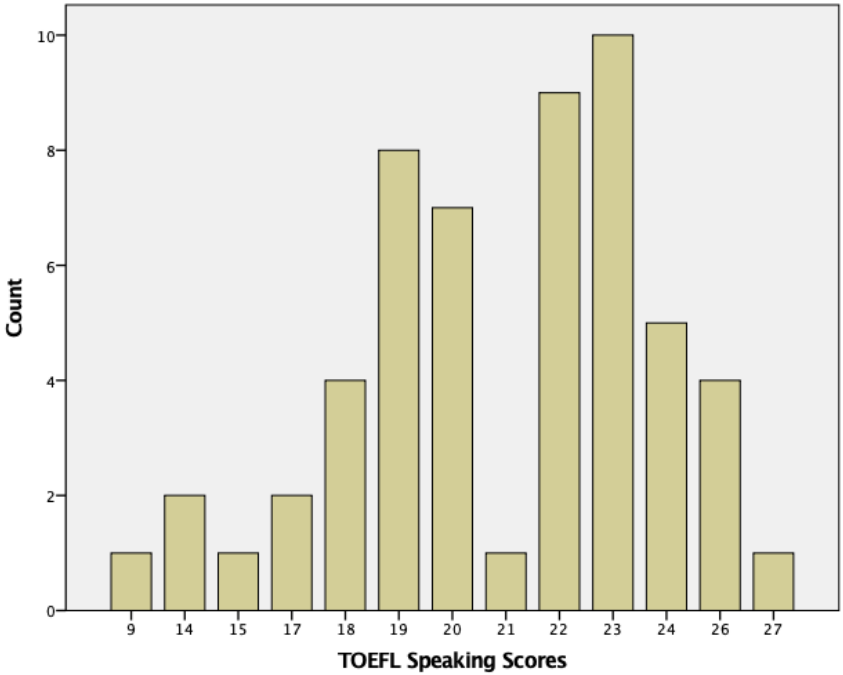


Figure 10: Distribution of TOEFL iBT® Speaking Scores

Figure 10 shows two peaks at 19 and 23 points with downwards slopes between them, on their left, and on their right.

Table 16: Distribution of TOEFL iBT® Speaking Scores

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
9	1	1.8	1.8
14	2	3.6	5.5
15	1	1.8	7.3

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
17	2	3.6	10.9
18	4	7.3	18.2
19	8	14.5	32.7
20	7	12.7	45.5
21	1	1.8	47.3
22	9	16.4	63.6
23	10	18.2	81.8
24	5	9.1	90.9
26	4	7.3	98.2
27	1	1.8	100.0
Total	55	100.0	

TOEFL iBT speaking scores of 26–30 are considered to be ***good***, 18–25 ***fair***, 10–17 ***limited***, and 0–9 ***weak*** (cf. Educational Testing Service, 2014). Table 14 shows that approximately 10 % of the participants were considered to be ***good*** speakers, 80 % were ***fair***, and 10 % were ***limited*** or ***weak***. Table 17 provides the descriptive statistics of the ACTFL and TOEFL iBT speaking results.

Table 17: Descriptive Statistics of ACTFL Speaking Proficiency Levels and TOEFL iBT® Speaking Scores

	ACTFL Levels	TOEFL iBT Score
Possible Range	1–10 (NL to S)	0–30
Observed Range	5–9 (IM to AH)	9–27
Median	7	22
Mean	7.16	20.91

	ACTFL Levels	TOEFL iBT Score
Standard Error of the Mean	0.15	0.46
Standard Deviation	1.09	3.37

Table 17 shows that both ACTFL and TOEFL iBT results involved a relatively narrow range (standard deviations are small). To align ACTFL ratings and TOEFL iBT scores, the numeric equivalencies of ACTFL sublevels were used. Figure 11 plots TOEFL iBT speaking scores and ACTFL speaking proficiency levels.

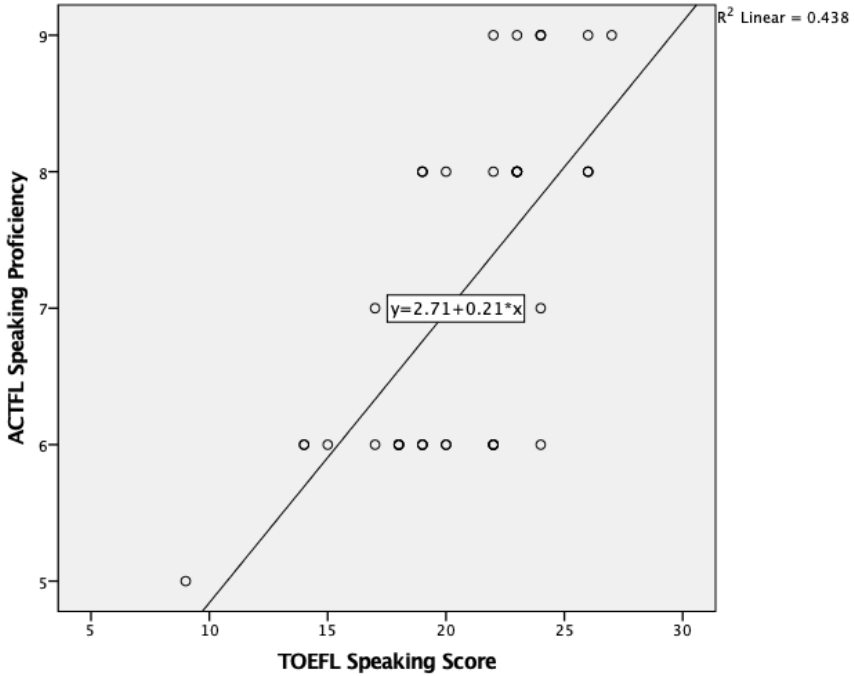


Figure 11: Scatter Plot of TOEFL iBT® Speaking Scores and ACTFL Speaking Proficiency Levels

Figure 11 shows the relationship between ACTFL speaking proficiency levels and TOEFL iBT speaking scores. TOEFL iBT speaking scores accounted for 43.8 % of the variance of the ACTFL proficiency level (R^2 = 0.438). This is a large effect. Effect sizes above R^2 = 0.25 are considered to be large. Figure 12 shows a P–P plot of the standardized residuals examining the assumption of normal distribution of the ACTFL and TOEFL iBT speaking data.

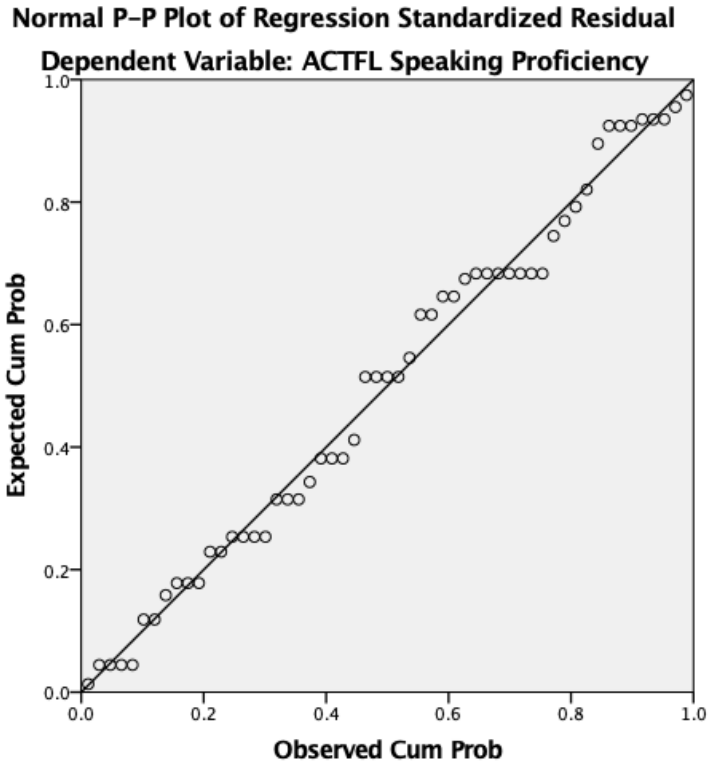


Figure 12: P–P Plot of the Standardized Residuals of TOEFL iBT® Speaking Scores and ACTFL Speaking Proficiency Levels

While there is some evidence of non-normality, the P-P plot by and large shows a linear relationship between TOEFL iBT speaking scores and ACTFL speaking proficiency levels. A linear regression analysis was used to predict ACTFL ratings on the basis of TOEFL iBT scores. The mean TOEFL iBT speaking score was $M = 20.91$, $S.E. = 0.455$; $SD = 3.373$, $N = 55$. The mean ACTFL speaking proficiency level was $M = 7.16$, $S.E. = 0.146$; $SD = 1.085$, $N = 55$. Pearson's correlation between TOEFL iBT speaking score and ACTFL speaking proficiency level was 0.662, $p < 0.01$ (2-tailed), $N = 55$. Both models explained 43.8 % of each other's results ($R^2 = 0.438$), which is a large effect. The maximum Cook's Distance was 0.089, supporting the assumption that there were no outliers.

The linear regression analysis with ACTFL rating as the dependent variable yielded a significant and large predictive effect of TOEFL iBT score on ACTFL rating: $p < 0.001$, Intercept (α): 2.713, Slope (β): 0.213. Table 18 shows the lowest TOEFL iBT speaking score predicting a particular ACTFL speaking proficiency level. Target ACTFL numeric values were whole numbers, i.e., the numbers associated with a particular ACTFL level. Accordingly, TOEFL iBT scores yielding numeric values closest to whole numbers were selected.

Table 18: Minimum TOEFL iBT® Speaking Scores Predicting ACTFL Speaking Proficiency Levels

ACTFL	IL	IM	IH	AL	AM	AH
ACTFL Numeric	3.99	5.06	5.91	6.97	8.04	9.10
TOEFL iBT	8	11	15	20	25	30

TOEFL iBT speaking scores of 8–10 predicted IL; scores of 11–14 predicted IM; 15–19 predicted IH; 20–24 predicted AL; 25–29 predicted AM; and 30 predicted AH.

Because the number of examinees was considerably smaller for speaking (and writing) than for the receptive skills, the median of the score ranges was generally used as the suggested cut score. The median slightly increases the number of false positives, but as Papageorgiou et al. (2015) argued, college admission decision makers are more concerned with reducing the number of false negatives. Setting the cut score at the lowest level increases the number of false negatives. Therefore, the median was used rather than the lowest score. In a few instances, the cut score was modified because of the score interpretations used by ETS (*good*, *fair*, *limited*, *weak*) (cf. Educational Testing Service, 2014) and the results of their standard setting studies. Table 19 shows the suggested correspondences between ACTFL reading proficiency levels and TOEFL iBT reading scores based on the present study. Because levels below ACTFL Intermediate are unlikely to be of interest to college admissions decision makers, the ACTFL Novice levels are excluded.

Table 19: Correspondences between ACTFL Speaking Proficiency Levels and TOEFL iBT® Speaking Scores

ACTFL	IL	IM	IH	AL	AM	AH
TOEFL iBT	9	13	18	22	26	30

For IL, the median of the IL range of 8–10 was used. ETS considers a score of 10 as ***limited*** speaking proficiency, which lends additional support to the suggested cut point, as well as the fact that their revised crosswalk (cf. Papageorgiou et al., 2015) considers a score of 10 to correspond to CEFR A2, which corresponds to ACTFL IL (cf. ACTFL, 2016). The suggested IL cut score of 9 is more conservative and should slightly decrease the number of false positives when compared with the more generous score of 10.

For IM, the median of the IM range of 11–14 is 12.5, rounded to 13. IM corresponds to CEFR B1, which according to the revised ETS crosswalk is associated with a TOEFL iBT score of 16, again making the suggested IM cut score more conservative. The median of the IH range of 15–19 is 17.5, rounded to 18. This cut score is also supported by the fact that ETS considers a cut score of 18 as a ***fair*** command of speaking proficiency. In addition, a cut score of 18 is the median of the B1 range of 16–19 established by the revised ETS crosswalk.

The median of the AL range of 20–24 is 22. AL corresponds to CEFR B2, which has a TOEFL iBT range of 20–24 according to the revised ETS crosswalk. In addition, it is more conservative than the minimally acceptable cut score of 23 established by Wylie and Tannenbaum (2006) for international teaching

assistants. The median of the AM range is 27.5, rounded to 28. However, because ETS considers cut scores of 26 and above to represent ***good*** levels of oral proficiency and because a cut score of 26 corresponds to a score of 50 on the Test of Spoken English (TSE), the more conservative value of 26 was selected as the equivalent of AM. For AH, the cut score of 30, which was the result of the regression analysis, was selected. AH corresponds to CEFR C1, which was associated with a TOEFL iBT score range of 28–30 in the original TOEFL iBT CEFR crosswalk and a score range of 25 to 30 in the revised one.

Writing Proficiency

A total of 58 participants had both TOEFL iBT writing scores and ACTFL writing proficiency ratings (WPT). Plotting a box plot of TOEFL iBT and ACTFL writing results revealed one outlier, which was removed, leaving a total of 57 participants for further analysis. Figure 13 and Table 20 present the distribution of ACTFL writing proficiency levels for the 57 participants.

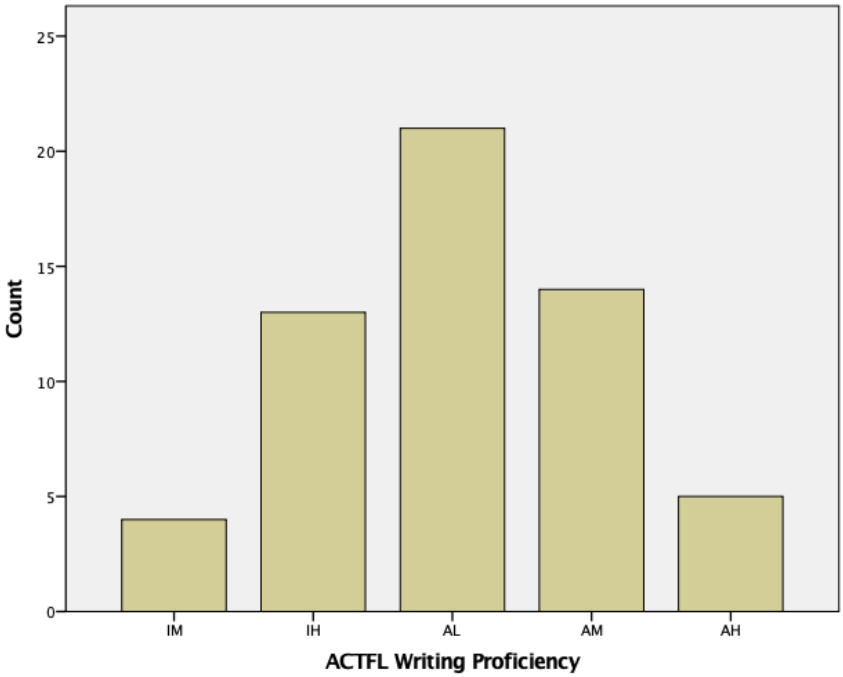


Figure 13: Distribution of ACTFL Writing Proficiency Levels
Note: IM = Intermediate Mid; IH = Intermediate High;
AL = Advanced Low; AM = Advanced Mid; AH = Advanced High

Figure 13 shows that the assessment results peaked at Advanced Low (AL) and sloped downwards to Intermediate High (IH) and Mid (IM) on the left and Advanced Mid (AM) and High (AH) on the right, exhibiting a close to normal distribution.

Table 20: Distribution of ACTFL Writing Proficiency Levels

ACTFL Numeric	ACTFL Level	Frequency	Percent	Cumulative Percent
5	IM	4	7	7
6	IH	13	22.8	29.8
7	AL	21	36.8	66.7
8	AM	14	24.6	91.2
9	AH	5	8.8	100
Total		57	100	

Table 20 shows that the largest number of participants were AL; approximately 30 % were Intermediate and close to 70 % were Advanced. Figure 14 and Table 21 show the distribution of the TOEFL iBT writing scores for the 57 participants who also had an ACTFL rating.

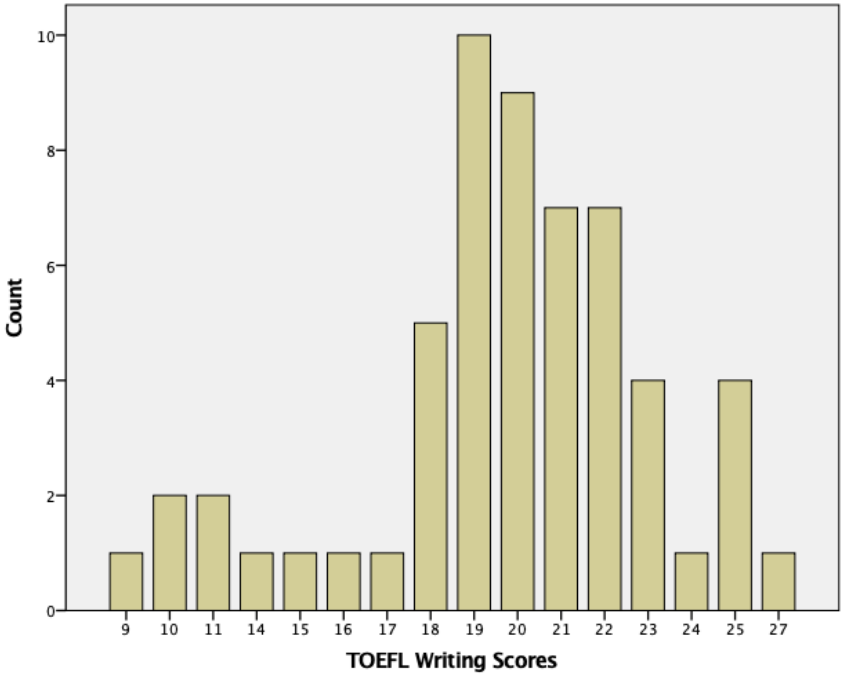


Figure 14: Distribution of TOEFL iBT® Writing Scores

Figure 14 shows a peak at 19 and two downwards slopes on the left and the right, indicating a relatively normal distribution.

Table 21: Distribution of TOEFL iBT® Writing Scores

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
9	1	1.8	1.8
10	2	3.5	5.3
11	2	3.5	8.8

TOEFL iBT Score	Frequency	Percent	Cumulative Percent
14	1	1.8	10.5
15	1	1.8	12.3
16	1	1.8	14.0
17	1	1.8	15.8
18	5	8.8	24.6
19	10	17.5	42.1
20	9	15.8	57.9
21	7	12.3	70.2
22	7	12.3	82.5
23	4	7.0	89.5
24	1	1.8	91.2
25	4	7.0	98.2
27	1	1.8	100.0
Total	57	100.0	

TOEFL iBT writing scores of 24–30 are considered ***good***, 17–23 ***fair***, and 1–16 ***limited*** (cf. Educational Testing Service, 2014). Table 21 shows that approximately 10 % of the participants were ***good***, 74 % were ***fair***, and 16 % were ***limited*** writers. Table 22 provides the descriptive statistics of the ACTFL and TOEFL iBT writing results.

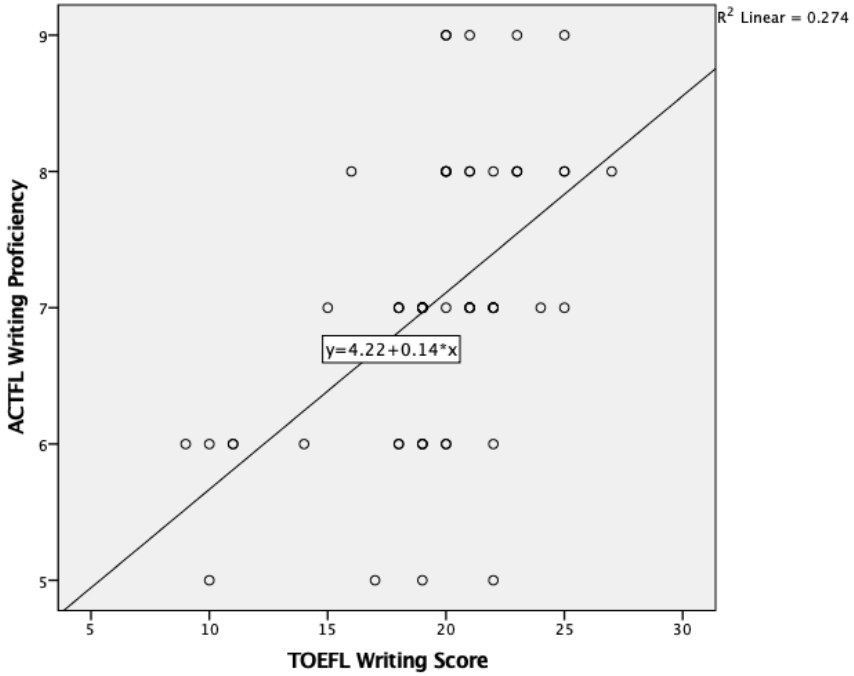


Figure 15: Scatter Plot of TOEFL iBT® Writing Scores and ACTFL Writing Proficiency Levels

Figure 15 shows the relationship between ACTFL writing proficiency levels and TOEFL iBT writing scores. TOEFL iBT writing score accounted for 27.4 % of the variance of the ACTFL writing proficiency level ($R^2 = 0.274$). This is a borderline large effect. Effect sizes above $R^2 = 0.25$ are considered to be large. Figure 16 shows a P–P plot of the standardized residuals examining the assumption of normal distribution of the ACTFL and TOEFL iBT writing data.

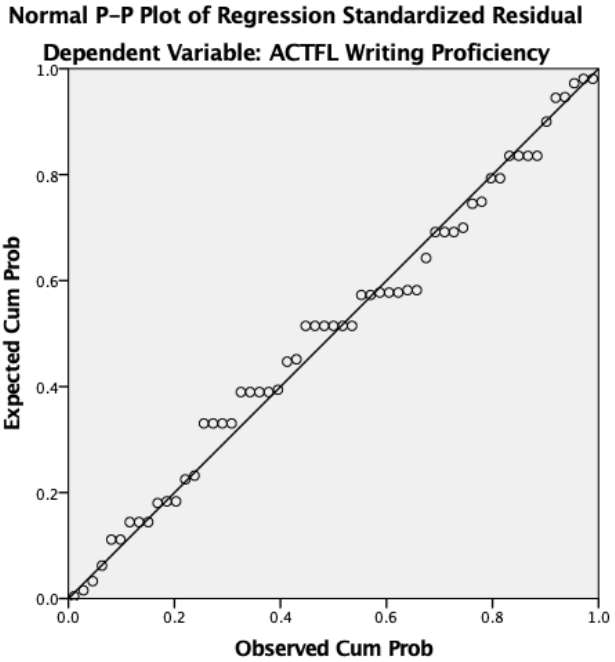


Figure 16: P-P Plot of the Standardized Residuals of TOEFL iBT® Writing Scores and ACTFL Writing Proficiency Levels

While there is some evidence of non-normality, the P-P plot by and large shows a linear relationship between TOEFL iBT writing scores and ACTFL writing proficiency levels. A linear regression analysis was used to predict ACTFL ratings on the basis of TOEFL iBT scores. The mean TOEFL iBT writing score was $M = 19.60$, *S.E.* = 0.509; $SD = 3.840$, $N = 57$. The mean ACTFL writing proficiency level was $M = 7.05$, *S.E.* = 0.140; $SD = 1.059$, $N = 57$. Pearson's correlation between TOEFL iBT writing score and ACTFL writing proficiency level was 0.523, $p < 0.01$ (2-

tailed), $N = 57$. Both models explained 27.4 % of each other's results ($R^2 = 0.274$), which is a borderline large effect. The maximum Cook's Distance was 0.089 supporting the assumption that there were no outliers.

The linear regression analysis with ACTFL rating as the dependent variable yielded a significant and borderline large predictive effect of TOEFL iBT score on ACTFL rating: $p < 0.001$, Intercept (α): 4.223, Slope (β): 0.144. Table 23 shows the lowest TOEFL iBT writing score predicting a particular ACTFL writing proficiency level. Target ACTFL numeric values were whole numbers, i.e., the numbers associated with a particular ACTFL level. Accordingly, TOEFL iBT scores yielding numeric values closest to whole numbers were selected.

Table 23: Minimum TOEFL iBT® Writing Scores Predicting ACTFL Writing Proficiency Levels

ACTFL	IL	IM	IH	AL	AM	AH
ACTFL Numeric	4.22	4.94	5.95	6.96	7.97	8.97
TOEFL iBT	1	5	12	19	26	(33)

TOEFL iBT writing scores of 1–4 predicted IL; writing scores of 5–11 predicted IM; 12–18 predicted IH; 19–25 predicted AL; and 26–30 predicted AM. The highest TOEFL iBT score is 30. Because only a non-existent score of 33 would have predicted AH, no TOEFL iBT score is selected to predict the AH level.

Because the number of examinees was considerably smaller for writing (and speaking) than for the receptive skills, the median of the score ranges was generally used as the suggested

cut score. The median slightly increases the number of false positives, but as Papageorgiou et al. (2015) argued, college admission decision makers are more concerned with reducing the number of false negatives. Setting the cut score at the lowest level increases the number of false negatives. Therefore, the median was used rather than the lowest score. In a few instances, the cut score was modified because of the score interpretations used by ETS (*good*, *fair*, *limited*, *weak*) (cf. Educational Testing Service, 2014) and the results of their standard setting studies. Table 24 shows the suggested correspondences between ACTFL writing proficiency levels and TOEFL iBT writing scores based on the present study. Because levels below ACTFL Intermediate are unlikely to be of interest to college admissions decision makers, the ACTFL Novice levels are excluded.

Table 24: Correspondences between ACTFL Writing Proficiency Levels and TOEFL iBT® Writing Scores

ACTFL	IL	IM	IH	AL	AM	AH
TOEFL iBT	3	8	17	22	26	30

For IL, the median of the IL range of 1–4 was used. ETS considers a score of 1 as *limited* writing proficiency, which lends additional support to the suggested cut point, as well as the fact that their revised crosswalk (cf. Papageorgiou et al., 2015) considers a score of 7 to correspond to CEFR A2, which corresponds to ACTFL IL (cf. ACTFL, 2016). The suggested IL cut score of 3 is fairly conservative, therefore, and should decrease

the number of false positives when compared with the more generous score of 7.

For IM, the median of the IM range of 5–11 is 8. IM corresponds to CEFR B1, which according to the revised ETS crosswalk is associated with a TOEFL iBT score of 13, again making the suggested IM cut score considerably more conservative. The median of the IH range of 12–18 is 15. However, because the texts produced by IH writers have much more in common with texts produced by AL writers than IM writers, the minimum TOEFL iBT score ETS considers ***fair*** proficiency is used, i.e., 17. This cut score is also supported by the fact that the first crosswalk, which ETS established on the basis of a standard setting, determined 17 as the lowest B1 score. According to the ACTFL CEFR crosswalk (cf. ACTFL, 2016), B1 corresponds to both IM and IH.

The median of the AL range of 19–25 is 22. AL corresponds to CEFR B2, which had a TOEFL iBT range of 21–27 in the original crosswalk and a range of 17–23 according to the revised ETS crosswalk. The cut score of 22 is part of both ranges, being at the lower end of the original crosswalk and the higher end of the revised crosswalk.

For AM, the lowest score of 26, which was established by the regression analysis, was selected as the cut score. ETS considers cut scores of 24 and above to represent ***good*** levels of writing proficiency. While the revised ETS crosswalk associates TOEFL iBT scores of 24–30 with C1, the original crosswalk associated TOEFL iBT scores of 28–30 with C1. The ACTFL CEFR

crosswalk associates AM with the upper half of B2. Therefore, the lowest score rather than the median was selected. For AH, a cut score of 30 was selected. While the regression analysis associated a score of 30 with AM, both the stricter original ETS crosswalk as well as the more lenient revised crosswalk associated a score of 30 with CEFR C1, which corresponds to AH.

Conclusion

The number of participants with reading and listening tests was sufficiently large (reading: $N = 195$; listening: $N = 197$) with high correlations (reading: $r = 0.796$; listening: $r = 0.708$) and effect sizes (reading: $R^2 = 0.634$; listening: $R^2 = 0.501$) to align ACTFL ratings and TOEFL iBT scores confidently. Using the categories established by ETS (***high***, ***intermediate***, and ***low***) for both reading and listening, some of the minimum TOEFL iBT values established empirically in this side-by-side study were adjusted slightly to better represent these TOEFL iBT categories.

The number of participants with speaking and writing tests was smaller (speaking: $N = 55$; writing: $N = 58$) with moderate correlations (speaking: $r = 0.662$; writing: $r = 0.523$) but still large effect sizes (speaking: $R^2 = 0.438$; writing: $R^2 = 0.274$). In addition, the range of the results was more restricted. A linear regression analysis was run with TOEFL iBT score as the predictor variable and ACTFL proficiency level as the dependent variable. The median of the ranges established by the regression analysis was used as the cut score except for AM and AH to minimize false negatives. As was the case with reading and listening, the final TOEFL iBT scores were adjusted to reflect the TOEFL iBT categories of ***good***, ***fair***, ***limited***, and ***weak*** more appropriately. Table 25 shows the recommended ACTFL TOEFL iBT correspondences based on the present study (cf. Tables 9, 14, 19, and 24). Superior (S) was added to Table 25

to account for S ratings as well. The recommended cut scores for S are the same as for AH.

Table 25: Recommended ACTFL TOEFL iBT® Correspondences

ACTFL	TOEFL iBT			
	Reading	Listening	Speaking	Writing
S	30	30	30	30
AH	30	30	30	30
AM	28	27	26	26
AL	22	22	22	22
IH	20	17	18	17
IM	15	14	13	8
IL	12	7	9	3

A few caveats should be considered. This study was based on participants from 9 colleges and universities, all in the U.S., and may not reflect the broader population of TOEFL iBT examinees. In addition, the majority of the students were graduate students (57 %) and most of them had Chinese as their first language (37 %). While the number of reading and listening tests administered was more robust (around 200 per skill), there were fewer speaking and writing tests (less than 60 per skill). Future studies should focus on increasing the number of speaking and writing tests as well as adding participants in other countries.

References

American Council on the Teaching of Foreign Languages (2016). ***Assigning CEFR Ratings to ACTFL Assessments***. Alexandria, VA: ACTFL. Online: https://www.actfl.org/resources/assigning-cefr-ratings-actfl-assessments, retrieved 7 February 2021.

Bärenfänger, O., & Tschirner, E. (2012). ***Assessing Evidence of Validity of Assigning CEFR Ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by Computer (OPIc).*** (Technical Report 2012-US-PUB-1). Leipzig: Institute for Test Research and Test Development.

Council of Europe (2001). ***Common European Framework of Reference: Learning, Teaching, Assessment***. Cambridge, UK: Cambridge University Press. Online: https://www.coe.int/en/web/common-european-framework-reference-languages/documents#Language_policy, retrieved 7 February 2021.

Council of Europe (2009). ***Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)***. Online: https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr, retrieved 7 February 2021.

Carlsen, C., & Deygers, B. (2014). ***The B2 Level and Its Applicability in University Entrance Tests***. Paper presented at the 5th ALTE International Conference, April 1, 2014, Paris, France. Online: https://lirias.kuleuven.be/1822355?limo=0, retrieved 7 February 2021.

Educational Testing Service (2014). ***A Guide to Understanding TOEFL iBT® Scores***. Princeton, NJ: Educational Testing Service. Online: https://docplayer.net/23893552-A-guide-to-understanding-toefl-ibt-scores.html, retrieved 7 February 2021.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting Performance Standards on Complex Educational Assessments. ***Applied Psychological Measurement***, 24, 355-366.

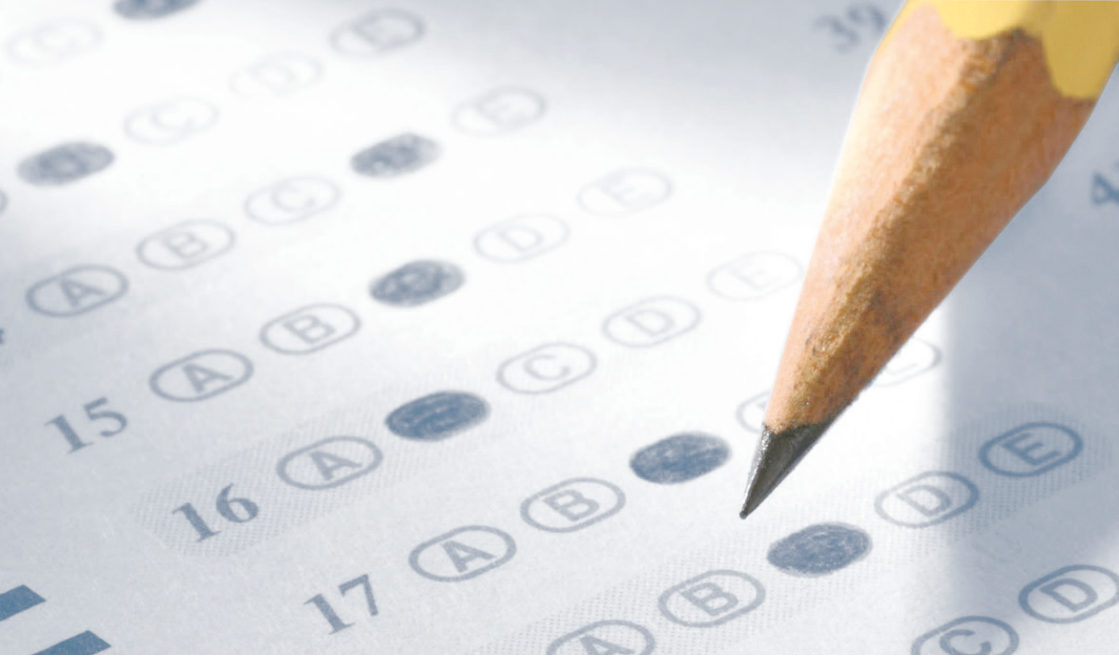
Impara, J. C., & Plake, B. S. (1997). Standard setting: An Alternative Approach. ***Journal of Educational Measurement***, 34, 353-366.

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 RM-15-06). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2005). ***Mapping English Language Proficiency Test Scores onto the Common European Framework of Reference.*** (Research Report RR-05-18). Princeton, NJ: Educational Testing Service.

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 No. iBT-06). Princeton, NJ: Educational Testing Service.

Wylie, E. C., & Tannenbaum, R. J. (2006). ***TOEFL iBT® Academic Speaking Test: Setting a Cut Score for International Teaching Assistants.*** (Research Memorandum RM-06-01). Princeton, NJ: Educational Testing Service.



INSTITUTE FOR TEST RESEARCH
AND TEST DEVELOPMENT