Measuring the Power of Learning.®

Research Memorandum
ETS RM–18-05

Best Practices for Comparing *TOEIC*® Speaking Test Scores to Other Assessments and Standards: A Score User's Guide

Jonathan Schmidgall

April 2018

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.

Best Practices for Comparing *TOEIC*® Speaking Test Scores to Other Assessments and Standards: A Score User's Guide

Jonathan Schmidgall
Educational Testing Service, Princeton, New Jersey

April 2018

Corresponding author: J. Schmidgall, E-mail: jschmidgall@ets.org

Suggested citation: Schmidgall, J. (2018). *Best practices for comparing* TOEIC® *speaking test scores to other assessments and standards: A score user's guide* (Research Memorandum No. RM-18-05). Princeton, NJ: Educational Testing Service.

Action Editor: Donald Powers

Reviewers: John Norris and Veronika Timpe-Laughlin

Abstract

In order to help score users understand the meaning of test scores, it may be helpful to relate scores on one test to performance levels on another relevant test or set of proficiency level standards. The process of demonstrating a relationship between two tests—or between a test and proficiency level standards—is typically referred to as *linking*, *concordance*, or *alignment*. A variety of research-based methods can be used to carry out this process. When score users rely on concordance tables that are not adequately backed by research, they are likely to make faulty inferences and inappropriate decisions about test takers. Arbitrary or inaccurate concordance may unfairly advantage some and disadvantage others while increasing the likelihood of decision errors. This memorandum provides a brief overview of best practices for (a) linking the scores of two tests and (b) establishing alignment between a test and a set of proficiency level standards. More specifically, it describes how the *TOEIC*® Speaking test and the American Council on the Teaching of Foreign Languages (ACTFL) Speaking standards have been aligned with the Common European Framework of Reference (CEFR) in order to enhance the meaning of test scores. Based on the research that supports these alignments, I hypothesize a relationship between TOEIC Speaking and ACTFL Oral Proficiency Interview computer-based (OPIc) test scores, which is subject to verification. Finally, I highlight a problematic example of a proposed linking of TOEIC Speaking and ACTFL OPIc test scores.

Key words: *TOEIC*®, concordance, linking, alignment, ACTFL, OPIc, CEFR

Important Things to Understand About Linking Scores From Different Tests

Typically, the meaning of scores is established, in large part, by a rigorous test development process in which the test is designed and scored on the basis of a clear and defensible description of the knowledge or skills to be assessed. The meaning of test scores can be further enhanced by aligning score levels with well-established proficiency standards or by linking the scores of two tests that measure the same or highly similar constructs (Kane, 2012). Alignment or concordance between tests (or between a test and standards) is typically summarized in a concordance table that shows how scores on one test correspond to scores on another.

Due to the desire to compare scores from different tests, score users have created concordance tables for a variety of tests, and they have proliferated online. Unfortunately, score users cannot necessarily take concordance tables at face value and instead need to distinguish between the good, the bad, and the ugly (Pommerich, 2007).

A few simple points should be considered when viewing concordance tables in order to determine their usefulness. Test score users should be highly skeptical of any concordance table that lacks documentation beyond the table itself. Without an adequate explanation of how the information in the table was produced, it is simply impossible to evaluate its trustworthiness. Determining that a concordance table is "ugly" requires knowing only that the table is not research-based.

Concordance tables that are based on research typically use one of four general methods, ranging from least rigorous to most rigorous: social moderation, prediction, scale alignment, or equating. The least rigorous approach, which is not based on statistical analyses, is referred to as *social moderation* (North, 2000). This approach is often used to align scales that employ qualitative descriptors, such as proficiency level standards, rather than scores. Due to expediency or a lack of quantitative data for analysis, alignment is justified through discussion and consensus between experts who are familiar with the tests or standards being aligned. The *prediction* approach uses a statistical approach to predict test takers' scores on one test based on their scores from another test. *Scale aligning* involves more complex research designs and statistical procedures that attempt to create a common scale from scores from two tests. The most rigorous approach is *equating*, which establishes a link between two tests so that scores can be

considered interchangeable (Holland & Dorans, 2006). Because this approach requires such strong assumptions—that the tests measure the same construct at the same level of intended difficulty and reliability—equating is typically used only for multiple forms of the same test rather than to link two different tests.

Whether a concordance table is deemed good or bad is determined mainly by the quality of the supporting research study and the claims made about scores (Pommerich, 2007). Many score users may find it difficult to evaluate the quality of research studies without additional expertise, but one general point should always be kept in mind: Regardless of the quality of the study, it is almost always incorrect to state that scores on one test are equivalent or equal to scores on another test without conducting an equating study. Research may show that experts believe that particular score levels on tests are related (i.e., social moderation approach), that a score on one test may predict a score on another with a stated degree of accuracy (i.e., prediction approach), or that the score scales may be linked with some confidence (i.e., scale aligning approach); however, unless an equating approach has been used, scores from different tests cannot be treated as equivalent or interchangeable—even if the test scores are highly or "significantly" correlated. A concordance table that implies that scores from different tests are equivalent without the use of equating is a clear reflection of bad practice, regardless of the quality of the research study that supports it.

Linking Tests to Standards: The *TOEIC*® Speaking Test, ACTFL OPI, and CEFR

A number of English language proficiency standards have been widely adopted and used by institutions and organizations globally. These standards enable policy makers, test designers, teachers, and learners to have a shared understanding of the categories used to define progressive levels of language proficiency. In addition, standards provide information about the language skills that language users are expected to possess and the communicative tasks they are expected to be able to perform at each level of proficiency. Two examples of widely used language proficiency standards include the Common European Framework of Reference (CEFR; Council of Europe, 2001) and the American Council on the Teaching of Foreign Languages (ACTFL; 2012) proficiency guidelines.

The CEFR Standards

The CEFR was created by the Council of Europe to facilitate a common understanding among Europeans regarding how to elaborate language proficiency across different levels in order to inform language learning, teaching, and assessment (for a brief overview, see Kantarcioglu & Papageorgiou, 2012). The CEFR was not designed with English or any other specific language in mind and has been published in more than 30 languages. Since its formal publication in 2001, its use has grown across the world and it has been used or adapted for non-European languages including Arabic and Japanese. It has a global community of users and its descriptive scheme has been actively researched since its initial publication, resulting in updates and additions (Council of Europe, 2018).

The CEFR standards provide descriptions of what language learners may be expected to do across three general proficiency levels: basic (A), independent (B), and proficient (C). Each general proficiency level is broken down into two sublevels, and the resulting classification of six proficiency levels is the most common use of the CEFR. The basic (A) level is divided into breakthrough (A1) and waystage (A2) sublevels, the independent (B) level into threshold (B1) and vantage (B2) sublevels, and the proficient (C) level into effective operational proficiency (C1) and mastery (C2) sublevels. These sublevels may be further divided (e.g., A2 into A2.1 and A2.2), and the descriptors used to characterize language ability at each level are prepared with a specific language competency in mind. Most commonly, the six sublevels are further divided through the use of "plus" levels between A2 and B1 (A2+), B1 and B2 (B1+), and B2 and C1 (B2+). The CEFR (Council of Europe, 2001, 2018) includes standards that have been developed for a wide variety of communicative language activities, strategies, and competencies, which are then articulated using more general scales (e.g., Overall Reading Comprehension) and more specific scales (e.g., Reading for Information and Argument).

In the CEFR standards, speaking ability is described using two general scales (Spoken Interaction, Spoken Production) and many specific subscales (e.g., Spoken Production: Public Announcements; Spoken Interaction: Information Exchange). The authors of the CEFR standards produced brief "can-do" descriptors for each general scale that can be used for self-assessment purposes (Council of Europe, 2001, pp. 26–27). Table 1 provides excerpts of some of these descriptors for the Spoken Production scale.

Table 1. Excerpts of Descriptors From the Spoken Production Category of the CEFR

CEFR level	Self-assessment descriptor
C2	"I can present a clear, smoothly flowing description or argument in a style appropriate to the context"
C1	"I can present clear, detailed descriptions of complex subjects integrating subthemes"
B2	"I can present clear, detailed descriptions on a wide range of subjects related to my field of interest"
B1	"I can connect phrases in a simple way in order to describe experience and events, my dreams, hopes, and ambitions"
A2	"I can use a series of phrases and sentences to describe in simple terms (familiar topics)"
A1	"I can use simple phrases and sentences to describe where I live and people I know"

Note. Adapted from *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* by Council of Europe, 2007, pp. 26–27. Copyright 2001 by Cambridge University Press.

The self-assessment descriptors shown in Table 1 provide a summary of the characteristics of a language user's spoken production at each CEFR level (see Council of Europe, 2001). These can-do descriptors focus on the communicative tasks that language users are able to perform at each proficiency level, not what they are unable to do. At the basic levels (A1 and A2), language users are expected to be able to produce limited speech (i.e., simple phrases and sentences) to describe immediate needs. At the independent levels (B1 and B2), speakers should be able to produce more coherent and extended speech on a wider variety of topics. At the proficient levels (C1 and C2), speakers should be able to produce clear, connected, and sustained discourse appropriate to a variety of professional or academic contexts.

The ACTFL Standards, OPI, and OPIc

The ACTFL proficiency guidelines were originally developed in 1986 to support the evaluation of functional language ability in language learning environments in the United States (ACTFL, 2012). Like the CEFR, the ACTFL guidelines were not developed specifically for English and are used for many different languages.

The ACTFL guidelines specify five general levels of proficiency: novice, intermediate, advanced, superior, and distinguished. Each of the first three levels (novice, intermediate, advanced) are further divided into low, mid, and high sublevels. The ACTFL proficiency guidelines do not specify further sublevels and are limited to each of the four functional language skills: reading, listening, speaking, and writing.

In the ACTFL (2012) proficiency guidelines, speaking ability is characterized as interpersonal or presentational. This characterization corresponds roughly to the distinction made

in the CEFR standards between spoken interaction and spoken production. Descriptors have been prepared for each of the five general levels (see Table 2) and all sublevels.

Table 2. Excerpts of Descriptors From the ACTFL Speaking Proficiency Guidelines

Proficiency level	Descriptor
Distinguished	Able to use language skillfully, and with accuracy, efficiency, and effectiveness . . . are educated and articulate users of the language (p. 4)
Superior	Able to communicate with accuracy and fluency in order to participate fully and effectively in conversations on a variety of topics (p. 5)
Advanced	Engage in conversation in a clearly participatory manner in order to communicate information (on a variety of topics) (p. 5)
Intermediate	When talking about familiar topics related to their everyday life . . . can ask simple questions . . . produce sentence-level language (p. 7)
Novice	Can communicate short messages on highly predictable, everyday topics (p. 9)

Note. Adapted from *ACTFL proficiency guidelines 2012*, by American Council on the Teaching of Foreign Languages, 2012, pp. 4–9. Copyright 2012 by ACTFL.

As shown in Table 2, the length, complexity, and fluency of speech produced by language users are expected to increase with proficiency level. As for the CEFR, speakers at the lowest levels (e.g., novice, intermediate) are generally expected to produce only limited speech (e.g., short messages) that is appropriate for familiar or predictable situations. Speakers at higher levels of proficiency (e.g., advanced, superior, distinguished) are expected to produce more extended speech appropriate for a range of contexts and communicative tasks.

The ACTFL Oral Proficiency Interview (OPI) and its computer-based version (OPIc) are speaking tests designed to produce evaluations of test takers aligned with the ACTFL Speaking proficiency guidelines. The ACTFL OPI is intended to test functional language ability or "what individuals can do with language in terms of speaking . . . in real-world situations in a spontaneous and non-rehearsed context" (ACTFL, 2012, p. 3). Although the same criticism can be made of the CEFR standards, experts have criticized the OPI and the ACTFL guidelines for creating a developmentally oriented scale (beginner, intermediate, advanced, superior) without any basis in language acquisition theory and limited external validation; without this evidence, many researchers continue to question what the ACTFL Speaking scale actually measures (see Malone & Montee, 2010, for critique of ACTFL; see Alderson, 2007, and Hulstijn, 2007, for a critique of CEFR).

The TOEIC Speaking Test, ACTFL Standards, and CEFR Standards

The *TOEIC*® Speaking test was designed to evaluate English speaking ability in the context of a global workplace and everyday life (Hines, 2010). Test takers complete 11 speaking

tasks that correspond to three underlying claims about what constitutes English speaking ability in this context: the ability to produce intelligible speech; to carry out routine social and occupational interactions; and to create connected, sustained discourse appropriate to the typical workplace. A minimum of three expert raters evaluate each test taker's response using scoring rubrics designed with the three underlying claims in mind (Everson & Hines, 2010). A scale score between 0 and 200 is assigned based on rater scores; the result is a reliable score (Schmidgall, 2017) that can be interpreted as a test taker's speaking ability in the global workplace and everyday life.

In order to help test takers and score users better understand the meaning of TOEIC Speaking test scores and provide more feedback, Educational Testing Service (ETS) researchers analyzed test-taker responses to determine the generalizations that could be made about test-taker abilities at different levels of the score scale. Ultimately, researchers identified eight proficiency levels (ETS, 2010). For each proficiency level, descriptions of the knowledge and skills typically exhibited by test takers at that level were generated (see the appendix).

The availability of details about TOEIC test-taker abilities at various score levels is believed to be a useful enhancement to TOEIC test score reports. For some score users, it may be more useful to interpret speaking ability with respect to levels specified in more widely used language proficiency standards. In order to support these score users, ETS research has related TOEIC Speaking test scores to a number of language proficiency standards, including CEFR (Tannenbaum & Wylie, 2008), STANAG 6001 (Tannenbaum & Baron, 2010), and standards for internationally trained nurses (Tannenbaum, 2010). Recently, TOEIC Speaking test score users have expressed interest in understanding the relationship between TOEIC Speaking test scores and ACTFL Speaking proficiency levels.

Although the relationship between TOEIC Speaking test scores and ACTFL proficiency levels has not been directly examined, both have been related to the CEFR by their sponsoring organizations. In a study of the relationship between TOEIC Speaking test scores and CEFR proficiency levels that included 22 experts representing 10 countries, researchers recommended minimum TOEIC Speaking score levels that convincingly differentiated test takers across the A1 to C1 proficiency levels (Tannenbaum & Wylie, 2008). The minimum scores (cut scores) used to indicate CEFR levels are shown on the left side of Figure 1. An arrow links each cut score on the TOEIC Speaking test score scale to a line that indicates the lowest point of each CEFR

proficiency level. As a form of the social moderation approach to concordance, the research study reflected experts' judgments about the score levels perceived to be associated with CEFR proficiency levels.

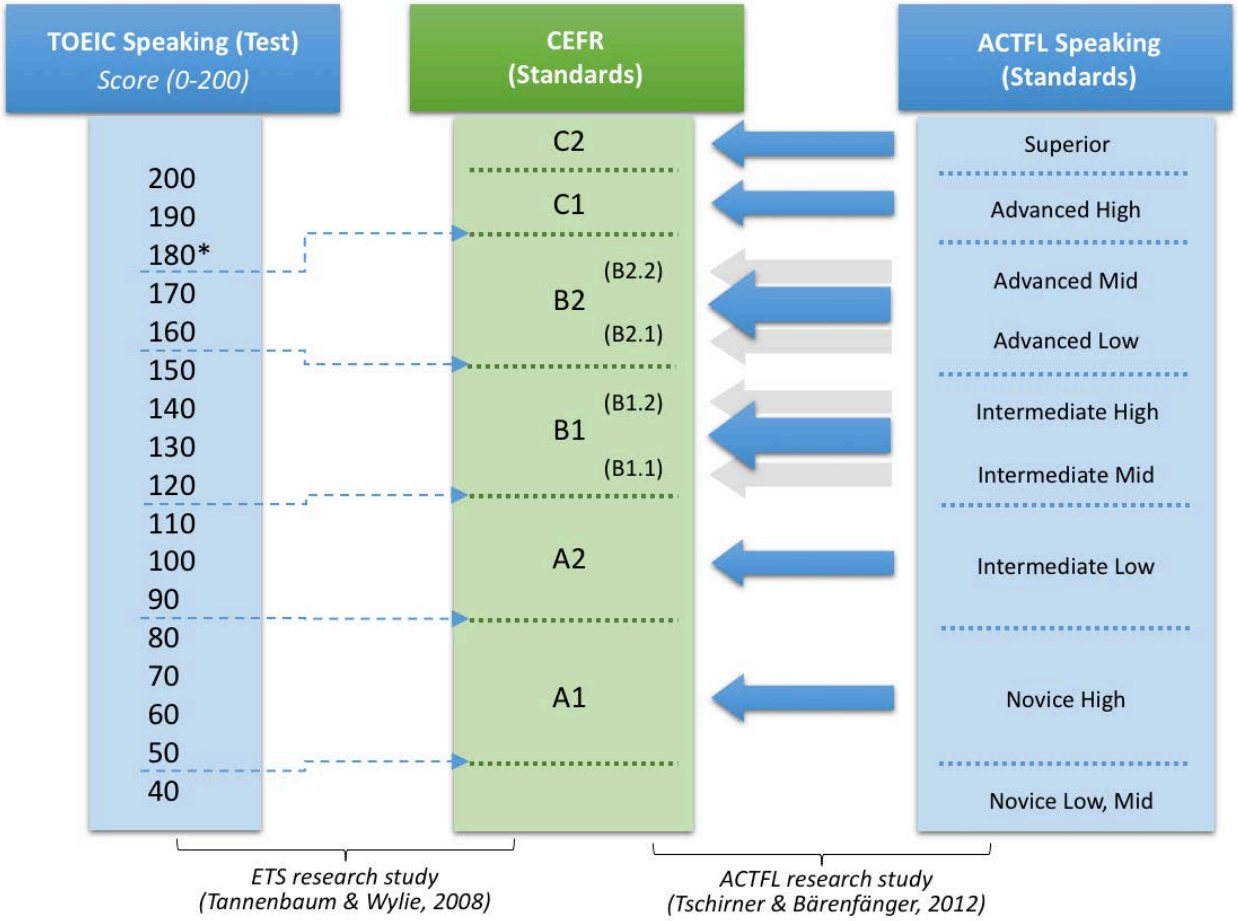

Figure 1. One-directional alignment between TOEIC speaking test scores and CEFR proficiency levels and between ACTFL and CEFR proficiency levels.

*The C1 cut score was adjusted from the recommended study values.

An alignment between ACTFL proficiency levels (as indicated by performance on the OPI and OPIc) and CEFR proficiency levels was established in a study by Tschirner and Bärenfänger (2012). Based on an analysis of the correspondence between ratings of CEFR and ACTFL (i.e., OPI and OPIc) proficiency levels for a sample of German speakers, the researchers proposed the correspondence between ACTFL and CEFR levels shown on the right side of Figure 1. As indicated by the arrows in Figure 1, performances assigned an ACTFL proficiency level rating of novice high corresponded to those assigned a CEFR proficiency level rating of

A1, and so forth. The light gray arrows that connect intermediate mid with CEFR B1.1, intermediate high with B1.2, advanced low with B2.1, and advanced mid with B2.2 indicate that these mappings were not empirically determined. In the study, raters used a scoring rubric that consisted of CEFR B1 and B2, not B1.1, B1.2, B2.1, and B2.2. After the researchers determined that ACTFL levels intermediate high and intermediate mid were judged to correspond to Level B1 of the CEFR, they suggested that intermediate high and intermediate mid may correspond to B1.1 and B1.2, respectively. The same logic was used to link advanced mid to B2.2 and advanced low to B2.1.

Figure 1 is described as a one-directional alignment because the research studies supported links from TOEIC Speaking test scores and ACTFL Speaking proficiency levels to CEFR levels, but not necessarily vice versa. As researchers have stated, CEFR proficiency levels are broad and vaguely defined; consequently, tests or test scores that have been aligned to the same CEFR levels should not be viewed as equivalent (North, 2014). While Figure 1 should not be used to assert equivalency between TOEIC Speaking test score levels and ACTFL Speaking proficiency levels, it provides a hypothesis for how they may be related through their common mapping to the CEFR. The ACTFL proficiency level advanced high has been mapped to CEFR Level C1, as have TOEIC Speaking test scores 180–200. ACTFL proficiency levels advanced mid and low have been mapped to CEFR Level B2, as have TOEIC Speaking test scores 160–170. ACTFL proficiency levels intermediate high and mid have been mapped to CEFR Level B1, as have TOEIC Speaking test scores 120–150. ACTFL proficiency levels intermediate low and novice high have been mapped to CEFR Levels A2 and A1, respectively, as have TOEIC Speaking test scores 90–110 and 50–80. More research is needed to directly support these hypothesized relationships, but the hypothesis itself is prompted by prior research.

An “Ugly” Example

Although a research-based concordance table between the TOEIC Speaking test and the ACTFL Speaking standards is still needed, score users may apply what they know about best practices in concordance tables and the initial expectations offered by Figure 1 to examine an existing concordance found online (http://blog.naver.com/PostView.nhn?blogId=pyohj21&logNo=220234487456) and reproduced in Table 3. This table intends to communicate a concordance between TOEIC Speaking test proficiency levels and OPIc proficiency levels based on the version of OPIc used in South Korea to assess English speaking proficiency.

Table 3. "Ugly" Concordance Table Between TOEIC Speaking and OPIc (ACTFL Speaking Proficiency Level)

TOEIC Speaking	OPIc
Level 8 (190–200)	AD (advanced)
Level 7 (160–180)	IH (intermediate high)
Level 6 (130–150)	IM (intermediate mid)
Level 5 (110–120)	IL (intermediate low)
Level 4 (80–100)	NH (novice high)
Level 3 (60–70)	NM (novice mid)
Level 1, 2 (0–50)	NL (novice low)

Several issues with this concordance table should cause immediate concern for a score user. First, there appears to be no research to support this particular table: It is presented without reference to any research study or additional documentation. Without this information, it is not possible to know whether the concordance table is good or bad; it's simply "ugly" (Pommerich, 2007).

Second, even if the concordance table shown in Table 3 is simply a hypothesis, it does not appear to be a particularly good one. It treats TOEIC Speaking proficiency levels as the basis for decision-making rather than as descriptors used to illustrate the meaning of scores. In the ACTFL OPI and OPIc Speaking tests, interpretations about test-taker abilities are expressed in terms of proficiency levels to facilitate decision-making. Although this may appear to be a subtle distinction, it reflects differences in test design that would need to be considered to facilitate alignment or concordance between test scores.

Finally, and perhaps most substantially, when compared to the hypothesis shown in Figure 1—based on ETS and ACTFL research—the concordance suggested by Table 3 appears to underestimate OPIc proficiency levels in relation to TOEIC Speaking test scores. As illustrated in Figure 1, ACTFL OPIc levels advanced mid and low have been shown to be related to CEFR Level B2. In a separate study, TOEIC Speaking scores at Level 7 were found to be related to CEFR Level B2. This leads to an initial hypothesis that TOEIC Speaking proficiency Level 7 may reflect a level of speaking proficiency comparable to ACTFL levels advanced mid and low. Surprisingly, Table 3 asserts that TOEIC Speaking proficiency Level 7 is comparable to the ACTFL level intermediate high. Further comparisons between Figure 1 and Table 3 highlight additional inconsistencies. Based on Figure 1, TOEIC Speaking proficiency Level 6 is probably closer to ACTFL Speaking proficiency level intermediate high. However, Table 3 implies TOEIC Speaking proficiency Level 6 is comparable to ACTFL level intermediate mid. Figure 1

shows how the ACTFL proficiency guidelines divide the advanced levels into advanced low, advanced intermediate, and advanced high, while Table 3 labels an ACTFL level as simply advanced, an inaccuracy. Although additional research needs to be conducted to investigate the hypothesis drawn from Figure 1, it seems more plausible than the concordance presented in Table 3.

Based on this review, the use of the concordance table shown in Table 3 would likely lead to inaccurate or unfair decisions. For example, a score user may determine that applicants for a job should have a level of English speaking ability comparable to a TOEIC Speaking test score of 140. Table 3 implies that this score would correspond to an OPIc score (i.e., ACTFL level) of intermediate mid, whereas Figure 1 suggests it would likely correspond to intermediate high. Test takers at the intermediate mid level would be less likely to meet the targeted threshold of speaking ability and thus selecting them would constitute false positive decision errors, decreasing the accuracy of decision-making. In addition, estimates of speaking ability based on OPIc scores would be inaccurately high in comparison to TOEIC Speaking, thereby compromising the fairness of the decision-making process.

Final Comments

Ultimately, the hypothesized relationship between TOEIC Speaking test score levels and ACTFL proficiency levels suggested by Figure 1 should be further evaluated in an empirical study. Language proficiency standards like the CEFR use brief, generalized descriptors that can be interpreted in different ways by different individuals and institutions (Fulcher, 2004) and are arguably most useful as heuristics (North, 2014). Expert analysis has suggested that there are important differences in how the CEFR and ACTFL standards (and assessments based on them) have been conceived with respect to how they (a) define language, (b) characterize language proficiency, and (c) describe language development (Chapelle, 2012). These various concerns underscore the need for a more direct analysis of the relationship between TOEIC Speaking scores and ACTFL (OPIc) Speaking proficiency levels in order to provide stronger support for their hypothesized relationship.

Using the information in this brief overview, TOEIC Speaking test score users should be able to identify untrustworthy (or so-called ugly) concordance tables. Any concordance table that is presented without supporting documentation using a research-based approach is probably untrustworthy. Even when research is used to support a concordance table, several key points

should be kept in mind to distinguish good and bad concordances: the type of approach used, the quality of the study, and whether its claims are appropriate. Ultimately, decisions about test takers are too important to rely on information from concordance tables that may be arbitrary or of low quality.

References

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

American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines 2012*. Alexandria, VA: Author.

Chapelle, C. A. (2012). Seeking solid theoretical ground for the ACTFL-CEFR crosswalk. In E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the Common European Framework of Reference for Languages* (pp. 35–48). Tubingen, Germany: Stauffenburg Verlag.

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

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors.* Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Educational Testing Service. (2010). *User guide: Speaking and writing*. Retrieved from https://www.ets.org/s/toeic/pdf/toeic_sw_score_user_guide.pdf

Everson, P., & Hines, S. (2010). How ETS scores the *TOEIC*® Speaking and Writing test responses. In D. Powers (Ed.), *TOEIC*® *compendium* (1st ed., pp. 8.1–8.9). Princeton, NJ: Educational Testing Service.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253–266.

Hines, S. (2010). Evidence-centered design: The *TOEIC*® Speaking and Writing tests. In D. Powers (Ed.), *TOEIC*® *compendium* (1st ed., pp. 7.1–7.31). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663–667.

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

Kantarcioglu, E., & Papageorgiou, S. (2012). The Common European Framework of Reference. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide*

to second language assessment (pp. 82–88). Cambridge, UK: Cambridge University Press.

Malone, M., & Montee, M. (2010). Oral proficiency assessment: Current approaches and applications. *Language and Linguistics Compass, 4*(10), 972–986.

North, B. (2000). Linking language assessments: An example in a low stakes context. *System, 28*, 555–577.

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*(2), 228–249.

Pommerich, M. (2007). Concordance: The good, the bad, and the ugly. In N. Dorans, M. Pommerich, & P. Holland (Eds.), *Linking and aligning scores and scales* (pp. 200–216). New York, NY: Springer.

Schmidgall, J. E. (2017). *The consistency of TOEIC® Speaking scores across ratings and tasks* (Research Report No. RR-17-46). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12178

Tannenbaum, R. J. (2010). *Setting standards on the TOEIC® Writing and Speaking assessments for internationally trained nurses* (Research Memorandum No. RM-10-15). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Baron, P. A. (2010). *Mapping TOEIC® test scores to the STANAG 6001 language proficiency levels* (Research Memorandum No. RM-10-11). 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. TOEFLiBT-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Tschirner, E., & Bärenfänger, O. (2012). *Assessing evidence of validity of assigning CEFR ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by computer (OPIc).* Retrieved from http://www.global8.or.jp/OPIc%20CEFR%20Study%20Final%20Report%20pdf.pdf

Appendix

TOEIC Speaking Test Proficiency Levels, Scale Score Ranges, and Descriptors

Proficiency level	Scale score range	Descriptor
8	190–200	Typically, test takers at Level 8 can create connected, sustained discourse appropriate to the typical workplace. When they express opinions or respond to complicated requests, their speech is highly intelligible. Their use of basic and complex grammar is good, and their use of vocabulary is accurate and precise. Test takers at Level 8 can also use spoken language to answer questions and give basic information. Their pronunciation, intonation, and stress are at all times highly intelligible.
7	160–180	Typically, test takers at Level 7 can create connected, sustained discourse appropriate to the typical workplace. They can express opinions and respond to complicated requests effectively. In extended responses, some of the following weaknesses may sometimes occur, but they do not interfere with the message: minor difficulties with pronunciation, intonation, or hesitation when creating language; some errors when using complex grammatical structures; and some imprecise vocabulary. Test takers at Level 7 can also use spoken language to answer questions and give basic information. When reading aloud, test takers at Level 7 are highly intelligible.
6	130–150	Typically, test takers at Level 6 are able to create a relevant response when asked to express an opinion or respond to a complicated request. However, at least part of the time, the reasons for or explanations of the opinion are unclear to a listener. This may be because of the following: unclear pronunciation or inappropriate intonation or stress when the speaker must create language; mistakes in grammar; and a limited range of vocabulary. Most of the time, test takers at Level 6 can answer questions and give basic information. However, sometimes their responses are difficult to understand or interpret. When reading aloud, test takers at Level 6 are intelligible.
5	110–120	Typically, test takers at Level 5 have limited success at expressing an opinion or responding to a complicated request. Responses include problems such as: language that is inaccurate, vague or repetitive; minimal or no awareness of audience; long pauses and frequent hesitations; limited expression of ideas and connections between ideas; and a limited range of vocabulary. Most of the time, test takers at Level 5 can answer questions and give basic information. However, sometimes their responses are difficult to understand or interpret. When reading aloud, test takers at Level 5 are generally intelligible. However, when creating language, their pronunciation, intonation, and stress may be inconsistent.

Proficiency level	Scale score range	Descriptor
4	80–100	Typically, test takers at Level 4 are unsuccessful when attempting to explain an opinion or respond to a complicated request. The response may be limited to a single sentence or part of a sentence. Other problems may include: severely limited language use; minimal or no awareness of audience; consistent pronunciation, stress, and intonation difficulties; long pauses and frequent hesitations; and severely limited vocabulary. Most of the time, test takers at Level 4 cannot answer questions or give basic information. When reading aloud, test takers at Level 4 vary in intelligibility. However, when they are creating language, speakers at Level 4 usually have problems with pronunciation, intonation, and stress. For more information, check the Pronunciation Levels and Intonation and Stress Levels.
3	60–70	Typically, test takers at Level 3 can with some difficulty state an opinion, but they cannot support the opinion. Any response to a complicated request is severely limited. Most of the time, test takers at Level 3 cannot answer questions and give basic information. Typically, test takers at Level 3 have insufficient vocabulary or grammar skills to create simple descriptions. When reading aloud, speakers at Level 3 may be difficult to understand. For more information, check the Pronunciation Levels and Intonation and Stress Levels.
2	40–50	Typically, test takers at Level 2 cannot state an opinion or support it. They either do not respond to complicated requests or the response is not at all relevant. In routine social and occupational interactions, such as answering questions and giving basic information, test takers at Level 2 are difficult to understand. When reading aloud, speakers at Level 2 may be difficult to understand. For more information, check the Pronunciation Levels and Intonation and Stress Levels.
1	0–30	Typically, test takers at Level 1 leave a significant part of the *TOEIC* Speaking Test unanswered. Test takers at Level 1 may not have the listening or reading skills in English necessary to understand the test directions and/or questions.

Note. Adapted from *User Guide: Speaking and Writing*, by Educational Testing Service, 2010, pp. 10–13. Copyright 2010 by Educational Testing Service.