Research Memorandum
ETS RM–23-07

The Effectiveness of the TOEFL® Essentials™ Test for Distinguishing English Proficiency Levels

John M. Norris
Jeremy Lee

August 2023

ETS Research Memorandum Series

The Effectiveness of the TOEFL® Essentials™ Test for Distinguishing English Proficiency Levels

John M. Norris
ETS Japan, Tokyo, Japan

Jeremy Lee
ETS, Princeton, New Jersey, United States

August 2023

Corresponding author: J. Lee, E-mail: JYLEE001@ets.org

Suggested citation: Norris, J. M., & Lee, J. (2023). *The effectiveness of the TOEFL® Essentials™ test for distinguishing English proficiency levels* (Research Memorandum No. RM-23-07). ETS.

Action Editor: Larry Davis

Reviewers: Jonathan Schmidgall and Spiros Papageorgiou

Abstract

The TOEFL® Essentials™ test, launched in 2021, is an assessment of English language proficiency for use in informing higher education admissions decisions as well as for other purposes that call for an overall estimation of a learner's English proficiency in daily life and academic settings. The test design combines innovative item types and multistage adaptive testing to emphasize both efficiency of measurement and effectiveness in estimating learners' proficiency levels, from beginning to advanced users of English. A basic assumption underlying the validity of interpretations and uses for the test is that scores will consistently distinguish English learners at varying levels of proficiency. To investigate this assumption, test performance data were collected from foreign language learners of English enrolled in university English language programs in Canada and the United States. Learners' English proficiency levels were estimated a priori based on course placements within the program curricula, and teacher judgments of learner proficiency levels were also collected. We found that test scores exhibited high reliability estimates and learners' scores were spread across a wide range of score bands. Comparisons of test scores between a priori groupings by low, medium, and high abilities indicated that the test consistently distinguished between the groups. While strong positive correlations were found between teacher ratings of learners' proficiencies and corresponding test scores, teachers estimated the proficiency levels of learners slightly higher than test scores in relation to an external language proficiency framework. Learners also exhibited generally uneven profiles of ability across the four skills tested, raising questions about the extent to which course placements can serve as a reliable criterion variable for representing holistic proficiency levels. Implications for the interpretation of TOEFL Essentials test scores and directions for future validity research are discussed.

Keywords English language proficiency testing, validity argument, evaluation inference, extrapolation inference, reliability, criterion-related validity

Acknowledgments

We would like to acknowledge the anonymous test takers and host language programs and in particular the program coordinators who were instrumental in recruiting learners and supporting their participation. We also greatly appreciate the contributions by Shoko Sasayama to initial planning for the study and Esther Park in helping to establish the research sites. We also appreciate feedback on this manuscript provided by Spiros Papageorgiou, Jonathan Schmidgall, and Larry Davis.

Table of Contents

It is a fundamental professional and ethical responsibility of test providers to ensure that their tests meet certain minimum standards of quality, fairness, and usefulness, most often encapsulated in the technical notions of validity and reliability of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). In particular, when a new test is developed and launched, it is incumbent on the test provider to evaluate various attributes and assumptions—based on how the test was designed—in relation to how well the test actually functions in measuring a specific type of knowledge or ability (often referred to as a construct) within a particular population of test takers under operational testing circumstances. In addition, the extent to which interpretations made on the basis of test scores are warranted, and the uses to which those scores are put are justifiable, should be subject to investigation.

Within language testing, contemporary best practice (e.g., Chapelle & Lee, 2021) emphasizes that these various dimensions of test quality, interpretation, and use should be specified in some detail during the design and development of a new test and subsequently investigated systematically to provide evidence in support of, or against, the basic assumptions surrounding the test. Kane’s (1992, 2013) argument-based approach to validity, the assessment use argument proposed by Bachman and Palmer (2010), and illustrative work by Chapelle (e.g., 2012) and colleagues (Chapelle et al., 2008; Chapelle & Voss, 2021), provide useful starting points, frameworks, and examples of the critical dimensions of language tests that call for evaluative attention as well as fitting methods for their investigation.

Kane’s argument-based approach, in particular, has come to serve as a widely adopted framework for guiding the development of comprehensive validity arguments for language tests. Summarizing considerably here, any language test provider should, arguably, spell out how their test can reasonably inform a specific set of intended interpretations, uses, and related consequences by providing a carefully crafted argument about the rationale underlying design, delivery, scoring, and score interpretation of the test. This kind of validity argument structure builds on detailed claims organized into hierarchical inferences including (a) how a particular knowledge or ability domain is reflected in the test design or blueprint, (b) how performances are elicited and scored, (c) how scores are assumed to generalize beyond the given test performance, (d) what aspects of a person's knowledge or ability the scores can actually explain or predict, and (e) what interpretations and decisions or actions are warranted on the basis of scores. It is in light of this series of well-defined claims that the empirical evaluation of the test should follow, with the specific claims investigated through appropriate and rigorous methods to provide evidence supporting them or refuting them.

While the argument-based approach provides a comprehensive framework that accounts for the critical aspects of test quality and use that call for evaluation, it has also been suggested that not all of the claims call for attention at the same time or to the same degree (e.g., Norris, 2008). Indeed, as tests are initially put into use, it may be that certain segments of the overall validity argument should be prioritized because they present the presumed conditions under which subsequent claims may come into play, or they address the most critical questions posed by test score users. For example, Norris (2008) demonstrated how several critical claims, related to the ability of a language program placement test to both produce reliable scores and distinguish among groups of students at different known proficiency levels, were prioritized by test users and evaluators over other validity claims about the construct represented by test scores and the consequences of placement decisions (the latter only feasible on the basis of evidence for the former).

Validity evaluation, then, is never a once-and-done endeavor; rather, it consists of a series of investigations that test the various claims that underscore the “argument” of a given language test, and it proceeds on the basis of those claims deemed most important at a given moment in the development, delivery, and use of a test. The current study reports on one focused investigation of a key claim underlying the validity argument of a newly developed English proficiency assessment, the TOEFL® Essentials™ test. In the following section, a brief outline of the test is provided as well as an overview of the validity argument and the specific claim(s) of interest for the current study.

The TOEFL Essentials Test

The TOEFL Essentials test is a four-skills test of English language proficiency, intended for use in informing admissions decisions, related to academic contexts where English is the primary medium of communication, as well as for other purposes that call for an overall estimation of a learner’s English proficiency in daily life and academic contexts. The test takes approximately 90 minutes to complete, and it was designed to estimate English proficiency across a very wide range of levels, including the six levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) and the 12 levels of the Canadian Language Benchmarks (CLB; Centre for Canadian Language Benchmarks, 2012). For the listening, reading, and writing sections, the test adopts a multistage adaptive design, with test takers first completing items of medium difficulty in each skill and then being routed to items of lower or higher difficulty to probe their specific language abilities. For speaking, the test design is linear, with all item types designed to be accessible to lower proficiency learners as well as amenable to the demonstration of advanced speaking abilities. The items on each section of the test present a combination of highly efficient tasks that quickly estimate learners’ proficiency levels, with more communicative tasks that tap into learners' abilities to engage in extended receptive and productive language use of the sort encountered in both daily life and academic settings (for more details, see Davis, Norris, et al., 2023; Papageorgiou et al., 2021).

The TOEFL Essentials test was launched in 2021 as a fully remote, human-proctored test that test takers complete on their own computer devices at home. Writing and speaking sections are scored centrally at ETS by a combination of human raters and automated scoring technologies. Total test and section (for listening, reading, writing, and speaking) scores are reported in the form of 12 band levels accompanied by CEFR level estimates. Test takers receive initial estimated listening and reading scores immediately upon completion of the test, and official score reports are issued to test takers and designated receiving institutions within approximately 6 days.

The TOEFL Essentials Validity Argument

The validity argument for the TOEFL Essentials test, as outlined in Papageorgiou et al. (2021), provides a comprehensive list of general inferences targeted by the test along with associated claims related to each inference, specific warrants needed to support each claim, and the types of evidence that could be used to provide backing for—or to contradict—these warrants. While all of the inferences bear scrutiny in evaluating test score interpretation and use, some assume priority over others in establishing a validity foundation upon which the scores of a relatively new test can be attributed meaning in the first place (Kane, 2013; Norris, 2008). Of specific interest for the current study is an inference that comes early in the argument chain, related to the notion of "evaluation"—here, in relation to the capacity of the test to reliably evaluate differences in learner ability—and specific claims directly derived from that inference in the form of “extrapolation” and “utilization” inferences. We elaborate these validity argument claims next.

The evaluation inference in general has to do with the extent to which the test can be effectively and securely administered, test performances can be systematically elicited and reliably scored, item and section scores can be calculated and combined into total test scores, and section and total scores can serve to accurately reflect test-takers’ language proficiency differences. The core evaluation claim in the TOEFL Essentials validity argument states, “Observations of performance on the TOEFL Essentials test tasks are evaluated to produce scores reflective of targeted language abilities” (Papageorgiou et al., 2021, p. 40). A key warrant, or supporting claim, and hence the focus of the current study, is that “test tasks distinguish among examinees with varying degrees of proficiency” (Papageorgiou et al., 2021, p. 41). Fundamentally, it is important to demonstrate that the test can be used consistently to identify learners with higher and lower levels of English proficiency, as reflected in score differences. The basic evidence for backing, or disputing, this claim consists of study findings that show how well the test section and total scores actually discriminate (i.e., indicate differences) among test takers with real English proficiency differences, as well as how reliably they do so.

A subsequent inference in the TOEFL Essentials validity argument is closely related to this specific evaluation inference and claim. The primary claim of the extrapolation inference highlights that “performance on the test is related to real-life measures of language proficiency within the context of use” (Papageorgiou et al., 2021, p. 43). Here, the assumption is that, if the evaluation claim of a reliable measurement that distinguishes among proficiency levels is supported, then other real-life indicators of test-takers’ proficiency differences should also be closely associated with test scores. A specific type of backing for this claim comes in the form of "test scores are associated with indicators of real life performance such as…teachers' judgements" (Papageorgiou et al., 2021, p. 43). Another directly related derivative claim is found in the utilization inference, where the fact that "test scores are mapped to external language proficiency levels" (Papageorgiou et al., 2021, p. 43) should enable test score users to make reasonable interpretations about test scores in relation to actual test takers.

In the current study, these three claims played a role in guiding the types of evidence gathered and the comparisons made on their basis. The study was designed to examine the reliability of test scores and their effectiveness at distinguishing among learners with known proficiency differences (evaluation inference), and it examined the relationship between test scores and teachers' judgments of learners' English language proficiency (extrapolation inference). Both teacher judgments and learners' known proficiency differences were determined on the basis of a common language proficiency framework (thereby invoking the utilization inference). This set of claims was also prioritized for investigation because of the close relationship between test score reliability, effectiveness of scores in distinguishing among proficiency levels, and teachers' intended uses of test scores for admitting and placing students into university English-language and English-medium programs of study.

Criterion-Related Validity Research

There are several possibilities for validity research that can produce evidence relevant to these kinds of claims for a language proficiency test. Important here is the assumption, firstly, that test-takers' performances not only are amenable to consistent scoring but also that those scores are consistently related to meaningful differences in proficiency. The consistency of scores is typically addressed through the estimation of score reliability, though it is important to demonstrate that acceptable levels of reliability hold across actual samples of test-taker populations under operational circumstances. Demonstrating the relationship of test scores to meaningful differences in language proficiency, then, calls for comparison with some kind of criterion measure that is trusted to reflect such differences. Furthermore, the incorporation of criterion measures that reflect real-life differences and that are grounded in well-known frameworks for describing language proficiency differences may play a useful role in investigating the interrelated string of claims running from evaluation to extrapolation and utilization.

The family of research associated with these types of measures and comparisons is typically referred to as criterion-related validity investigations (Brown, 2005). As the name implies, meaningful criteria are sought out and compared with test scores to determine the extent to which the test is capable of making distinctions among test takers in ways and to degrees similar with the criterion measures. One useful criterion takes the form of groups of test takers with known levels of proficiency in the ability being tested, and the corresponding investigation compares test scores for these known groups to determine whether differences are detected. For example, Norris (2008) compared performances on a new placement test by students who had advanced into the different levels of a 4-year German language curriculum to determine how well test scores were able to distinguish among each curricular level/year. Another common criterion measure for educational assessments has to do with expert or teacher judgments of learners’ abilities, which are compared with the learners’ corresponding scores on the test under investigation. For example, Papageorgiou and Cho (2014) compared teacher judgments of English learners’ proficiency levels with their scores on the TOEFL Junior® test.

For both of these approaches, it is often useful to have a frame of reference for determining the test-takers' or learners' proficiency levels in order to form groups that are meaningfully different or to provide familiar categories that teachers may use to make judgments of students' abilities. Such frames of reference may take several forms. A language program or curriculum consisting of courses that correspond to different known levels of learner ability is one common source for differentiating learner differences. Another possibility comes in the form of well-known language proficiency frameworks that provide meaningful descriptions of distinct levels, such as the CEFR or CLB. Reference to such common language proficiency frameworks can prove particularly useful in providing a middle ground where diverse phenomena-like test scores, program/curriculum levels, and teacher judgments can be brought together on a common scale. For example, Fleckenstein et al. (2018) investigated the extent to which teachers' CEFR-based judgments of English learners' language achievement in high school proved comparable to test-based estimations of the same learners' CEFR levels (using the TOEFL ITP® test).

For language proficiency tests intended to be used in relation to educational settings, criterion-related validity investigations that draw on sources such as curriculum/program levels, teacher judgments, and language proficiency frameworks should provide insights into claims regarding the effectiveness of the test at distinguishing among proficiency levels that are meaningful to score users. The current study evaluated related claims for the TOEFL Essentials test on the basis of these criterion-related sources.

Research Questions

The following research questions guided the study:

RQ1: To what extent do total and section score distributions on the TOEFL Essentials test reflect a broad range of proficiency differences within the participant sample? (evaluation inference)

RQ2: How reliable are TOEFL Essentials test scores in an operational setting, in the form of Cronbach's alpha internal consistency estimates? (evaluation inference)

RQ3: To what extent do average TOEFL Essentials total and section scores distinguish among learner proficiency groups as determined by their placement into English program courses? (evaluation inference, extrapolation inference)

RQ4: What is the strength of association between TOEFL Essentials total scores and English program teacher judgments of learners' proficiency levels as aligned to a language proficiency framework? (extrapolation inference, utilization inference)

Methods

To address these research questions, data were collected from international students studying English as a second language at several university-level programs in Canada and the United States. All students completed official administrations of the TOEFL Essentials test, and criterion variables included (a) their placement into English language courses that were leveled according to either the CLBs or the Common CEFR levels and (b) teacher judgments of learner proficiency levels. Details on all aspects of the methodology follow.

Participants

A total sample of 143 participants completed all components of the TOEFL Essentials test and constituted the data set for the current study. College-level English learners were recruited for this study at two institutions in Canada (Institution A and Institution B) and one institution in the United States (Institution C). Note that students from the three institutions were combined into a single participant sample for the purposes of this study, primarily to have sufficient numbers of participants for meaningful statistical analyses. The pooling of participants was possible due to the proficiency-based organization of courses within each of the three programs. That is, each program referred to either the CLB or the CEFR in distinguishing the levels of their courses as well as the corresponding proficiency levels of students. Because the CLB and CEFR have also been mapped to each other (North & Piccardo, 2018), this course leveling allowed for courses and students at each institution to be compared with each other on the basis of assumed proficiency differences. Of course, it is possible that the interpretation and implementation of proficiency frameworks and levels differed across the three programs, but at a minimum their common organization of course levels according to these frameworks did provide one shared basis for comparison. In addition, it is possible that institution-specific analyses might have revealed unique patterns in the reliability and effectiveness of TOEFL Essentials test scores in relation to each program; however, individual program sample sizes were not sufficient for this additional level of analysis. In the following, we describe the proficiency-based structure of each program at each institution.

Institution A is a college in Canada that provides prematriculation English as a second language (ESL) classes at three different program/course levels that are aligned to CLB Levels 5 through 8 (with some overlap in students' CLB proficiencies possible at each of the three course levels). Courses at each level cover a variety of English learning foci. Students can be placed into a course level by taking an in-house placement test by submitting scores on recognized English proficiency tests or through pathway programs with other institutional partners. Students can also advance into the next course level if they received 70% or higher course grades in the previous term.

Institution B is a language school with several campuses in Canada. The courses at this institution are aligned with both CLB and CEFR levels. Institution B categorizes students into five proficiency levels that correspond to different available course levels: beginner (CLB Levels 1 and 2 and CEFR A1), preintermediate (CLB Levels 3 and 4 and CEFR A2), intermediate (CLB Levels 5 and 6 and CEFR B1), upper intermediate (CLB Levels 7 and 8 and CEFR B2), and advanced (CLB Levels 9 and 10 and CEFR C1). The institution uses an in-house placement test that includes four sections: grammar, reading, writing, and speaking. Overall scores on the placement test are used to place students in one of the five levels.

Institution C is a university located in the United States with an ESL program that supports international students. This program offers preparation in general academic English, with courses spanning eight levels aligned to CEFR levels A2 to C1. Like institutions A and B, students' placement in courses is determined by an in-house placement test. This test includes sections in reading, writing, speaking, and listening.

Participants were recruited by research partners at each institution, through email announcements and classroom visits that explained the study. Participation in this study was entirely voluntary and was not related in any way to students' classroom performance or course grades. Participants were encouraged to participate on the basis of receiving an official TOEFL Essentials test score, which they could use for a variety of approved purposes. The participants who agreed to participate in this study represented a variety of first language backgrounds, including Arabic, Chinese, Japanese, Korean, Portuguese, Russian, Spanish, and Vietnamese. Additional demographic data for participants were not collected in the current study to limit the intrusiveness of data collection for program partners; in addition, the singular demographic focus of meaning in the current study was participants' estimated English proficiency levels. Participants' estimated proficiency was determined by their current course enrollments, which were all aligned to either the CLB or CEFR levels, as well as independently through teachers' judgements. Information about participants' course enrollments was provided by the research partners at each institution (see the Instruments and Procedures sections in this report for further explanation on the assignment of participants to proficiency levels).

Instruments

All participants took the TOEFL Essentials test (see the prior detailed description). Several efforts were made to assist participants in registering for and taking the test, including the provision of (a) a document detailing the registration process in participants' first languages; (b) another document outlining the requirements for taking the test, such as a proper Internet browser, microphone, and webcam; (c) a walk-through video of the registration process; and (d) a study information sheet that provided basic information about the rationale and conduct of the study.

In order to collect information with which to independently estimate the English proficiency of participants, research partners at each institution reported the most recently enrolled courses for each participant as well as the corresponding estimated level of proficiency for each course in the form of either CLB or CEFR levels. For each participant, a teacher at their institution was assigned to provide a CLB or CEFR level estimate based on their knowledge of and interactions with the participants in their classes. Estimates were based on their most recent interactions in class, their assignments, and any relevant assessment scores. A single holistic proficiency estimate was provided for all participants as well as a skill-specific proficiency level estimate whenever applicable (e.g., if students were enrolled in a speaking and listening course, they would also receive proficiency estimates for those specific skills).

Procedures

Students who expressed interest in participating in the study received a voucher code that they then used to register for the operational TOEFL Essentials test at no cost. Note that various online resources were available at the TOEFL website, which the participants could use to familiarize themselves with the test format and content prior to testing, if desired. Participants were allowed to take the TOEFL Essentials test from their home or from the computer lab of their institution in keeping with current remote-testing protocols. All participants completed the test alone in a room that met security specifications, as verified by a remote proctoring service. Participants were not allowed to have any items on their desk and were required to show the proctor their surroundings before beginning the test. Participants completed the test entirely on their own, except for interactions with the remote proctor in cases of technical difficulties. For a few instances when technical difficulties caused testing to be suspended, participants reported it to the institutional research partner, and they were given another opportunity to take the test within 1 weeks' time. After completing the test, all participants received an official score report, which could be used as evidence of English language proficiency as with other TOEFL Essentials test takers.

The data collection took place throughout 2022 with participation windows made available by each institution a few weeks after the start of their academic term. All participants completed the TOEFL Essentials test within 2 to 4 weeks from the beginning of a given semester of instruction. The timing of test completion was controlled in this way to ensure that course level and teacher estimates of participants' proficiency were as accurate as possible and anchored at the point in time when the student commenced study at a given course level. The initial weeks of the semester also allowed teachers to develop a general sense of students' current English proficiency.

Scoring and Analyses

Participants' performances on the TOEFL Essentials test were officially scored at ETS following standard scoring protocols, and participants received an official score report. The score report included band scores for the four skill sections as well as a total band score, each ranging from 1 to 12 points. These band scores provided the basic data set for the current study.

To determine whether the TOEFL Essentials test effectively distinguishes among different levels of English proficiency, the participants were grouped by their course enrollment levels. Participants who were enrolled in courses aligned to CLB Level 5 (or CEFR A2) or below were placed in the "beginner" group. Participants in courses aligned to CLB Levels 6 or 7 (or CEFR B1) were placed in the "intermediate" group. Finally, participants in courses aligned to CLB 8 (or CEFR B2) and above were placed in the "advanced" group. It is important to point out that these groupings are broadly reflective of the proficiency levels typically represented in English language support programs at North American universities. Thus, a majority of students in such programs are at the CEFR B1 to B2 levels, where decisions regarding admissions tend to fall. Students with more advanced proficiency levels (C1 and C2) tend to be matriculated in regular courses of study and are generally less represented in English support programs, while students at the lowest CEFR levels (A1 and A2) are fewer in number because of their incipient language abilities and a lack of fitting course options. Our use of the terms "beginner," "intermediate," and "advanced" is intended to reflect this reality—where the majority of distinctions tend to be made between CEFR Levels A2, B1, and B2—and is not intended to align with the use of those terms by any proficiency framework. Based on this grouping strategy, 30 participants were in the beginner group, 63 were in the intermediate group, and 50 were in the advanced group. Equivalence between these ability categories and test scores were based on score mapping studies (Davis, Garcia Gomez, et al., 2023; Papageorgiou et al., 2021, 2022), which established the relationship between TOEFL Essentials band scores and CLB or CEFR levels (see Tables 1 and 2 for the score mappings on each scale).

Table 1. Mapping of TOEFL Essentials Test Scores to the Canadian Language Benchmarks (CLB) Levels

CLB	Speaking band score	Writing band score	Listening band score	Reading band score	Overall band score
11	12	12	12	12	12
10	11	11	11	11	11–11.5
9	10	9–10	10	10	10–10.5
8	7–9	7–8	9	8–9	8–9.5
7	6	6	7–8	7	6.5–7.5
6	5	5	6	6	5.5–6.0
5	4	4	5	4–5	4.5–5
4	3	3	3–4	3	3–4
3	2	2	2	2	2–2.5
2	1	1	N/A	N/A	N/A

Table 2. Mapping of TOEFL Essentials Test Scores to the Common European Framework of Reference for Languages (CEFR) Levels

CEFR level	Section band score (1–12)	Overall band score (1-12)
C2	12	12
C1	10–11	10–11.5
B2	8–9	8–9.5
B1	5–7	5–7.5
A2	3–4	3–4.5
A1	2	2–2.5
Below A1	1	1–1.5

Analyses took the form of descriptive statistics and graphical representation of score distributions for total and section scores, and test score reliability was estimated with Cronbach's alpha. Inferential statistics based on a univariate analysis of variance (ANOVA) were also calculated to examine possible score differences between the proficiency groupings, and group means and 95% confidence intervals were plotted, both for the section and total scores. Note that an experiment-wise alpha level was set at $\alpha < .05$, and a Bonferroni adjustment was made for the five inferential tests, reducing the critical alpha level for each ANOVA to $\alpha < .01$. Teacher proficiency estimates were compared with test scores using Spearman's rho correlation and Wilcoxon signed-rank test to account for the ordinal nature of the teacher estimates.

Results

The total sample of participants' test score data was first examined to determine the extent to which the TOEFL Essentials test was able to spread learners out into higher and lower scores according to the anticipated range of learner English proficiency. To this end, descriptive statistics were calculated for total and section scores overall, and histograms were plotted to depict the score distributions. Descriptive statistics were also calculated for performances by learners according to their a priori proficiency grouping into beginner, intermediate, and advanced groups (i.e., based on their course level enrollments and corresponding proficiency estimations for each course).

Figures 1–5 show the distributions of participants' TOEFL Essentials band scores for the total and section scores, and Table 3 provides descriptive statistics for the full participant sample as well as for each of the three proficiency groups. It is clear from the distributions and descriptive statistics that the test spreads learners out into a broad range of scores, with band scores extending from near the bottom to near the top of the available 12-point scale. For total, reading, and listening scores, the distributions assumed relatively normal curves, if slightly positively skewed for each, and with mean scores at or just under the mid-point of the scale. Speaking and writing scores were more positively skewed, with a majority of learners scoring on the lower to middle portion of the score scale. It is clear from these initial analyses that learners in this sample exhibited uneven English proficiency profiles, in the form of generally stronger receptive skills compared with their productive skills.

Descriptive statistics also show that the total and section scores exhibited expected variability, with standard deviation values around 2 to 2.5 points. Across scores for all of the proficiency groups, the listening scores were highest on average, followed by reading then speaking, with writing scores noticeably lower. Average scores on each of the sections, and the total, also differed in predicted patterns, consistently lowest for the beginner group, followed by the intermediate group, and highest across the four sections and total scores for the advanced group.

Figure 1. Frequency Distribution of TOEFL Essentials Total Scores

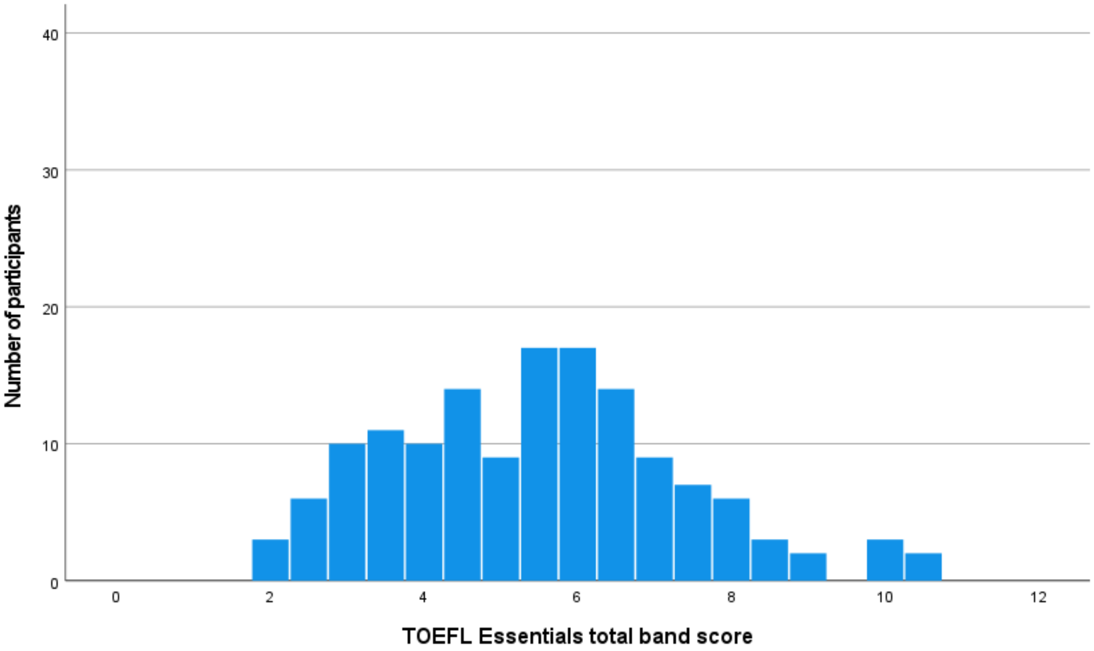


Figure 2. Frequency Distribution of TOEFL Essentials Reading Scores

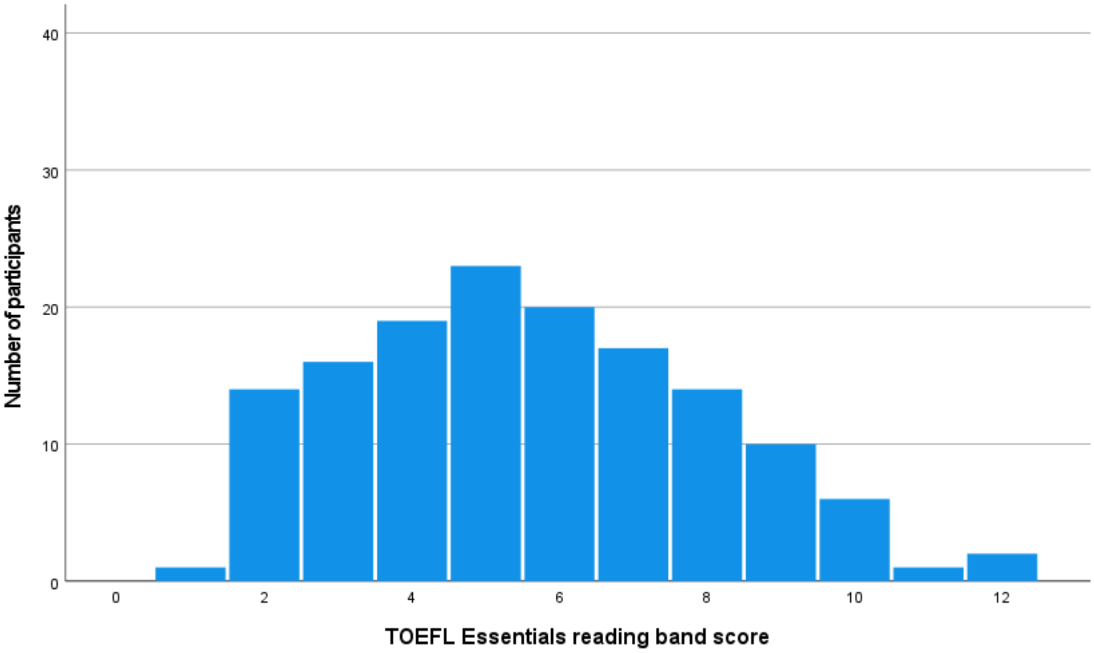


Figure 3. Frequency Distribution of TOEFL Essentials Listening Scores

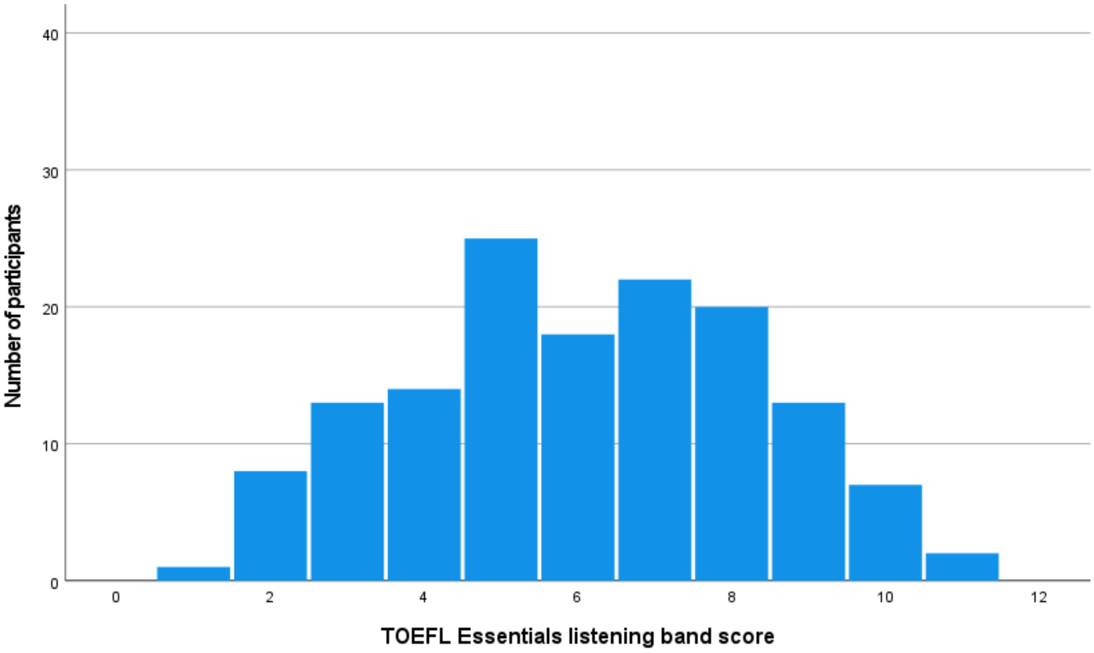


Figure 4. Frequency Distribution of TOEFL Essentials Speaking Scores

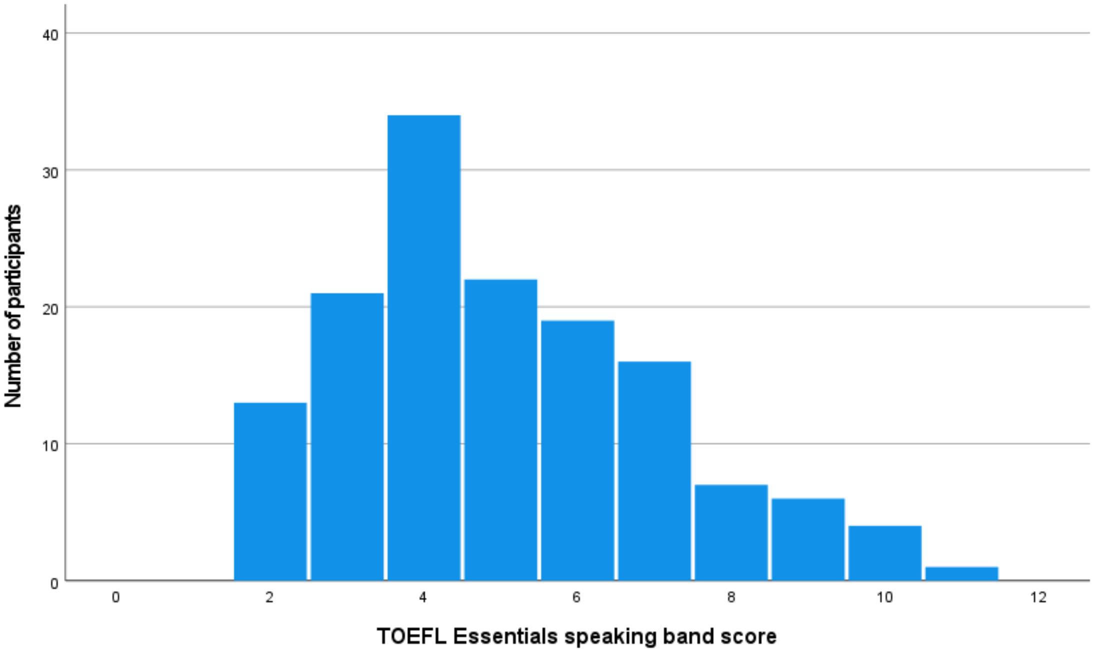


Figure 5. Frequency Distribution of TOEFL Essentials Writing Scores

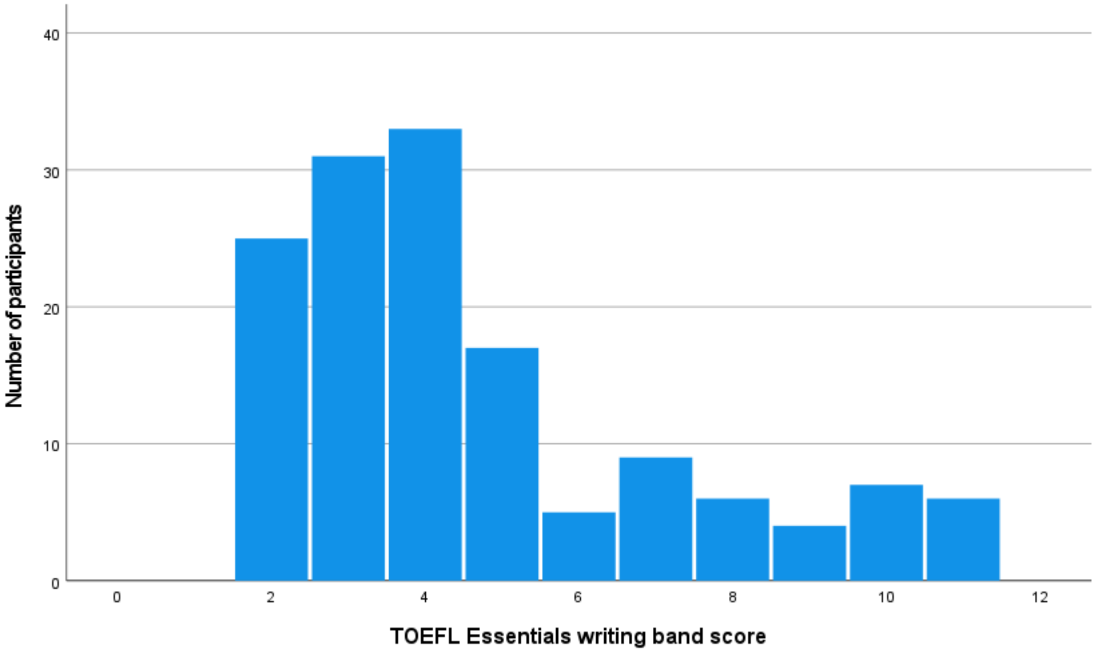


Table 3. Descriptive Statistics for TOEFL Essentials Total and Section Scores Overall and by Learner Proficiency Groups

Group	Section	*N*	Min	Max	Mean	*SD*
Overall	Reading	143	1	12	5.62	2.41
	Listening	143	1	12	6.07	2.27
	Speaking	143	2	11	5.05	2.08
	Writing	143	2	11	4.71	2.55
	Total	143	2	11	5.47	1.88
Beginner	Reading	30	1	8	3.67	1.81
	Listening	30	1	8	3.83	1.72
	Speaking	30	2	6	3.23	1.17
	Writing	30	2	10	2.87	1.50
	Total	30	2	6	3.52	1.09
Intermediate	Reading	63	2	11	5.67	2.23
	Listening	63	2	11	6.25	1.99
	Speaking	63	2	10	5.13	1.76
	Writing	63	2	11	4.52	2.27
	Total	63	3	10	5.50	1.45
Advanced	Reading	50	2	12	6.74	2.25
	Listening	50	3	11	7.18	1.95
	Speaking	50	3	11	6.04	2.19
	Writing	50	2	11	6.04	2.64
	Total	50	3	11	6.60	1.80

In order to examine the consistency with which total scores could be produced on the TOEFL Essentials test for this small sample of test takers, Cronbach's alpha was calculated for the total score based on the contributions of the individual section scores, and the resulting estimate revealed quite high reliability, α = 0.89. The high reliability estimate is also noteworthy, given that it was calculated on the basis of the four skill-section scores only, essentially treating the test as consisting of four polytomous items (one for each section score). This finding indicates that, even when estimated based on performances by a small and relatively idiosyncratic sample of test takers, the test scores exhibited an anticipated level of quite high-reliability.[1]

In order to probe further the extent to which the TOEFL Essentials test was able to distinguish among learners at different proficiency levels, performances by the three proficiency groupings were compared. It is clear from the descriptive statistics for each group that the average total test scores differed substantially, by 2 band level points between the beginner and intermediate groups, and by 1 band level point between the intermediate and advanced groups. A univariate ANOVA ($F(140, 2) = 38.57$, $p = 0.000$) indicated statistically significant differences for proficiency groupings on the TOEFL Essentials total band score, and these differences between the means for each group are clear in Figure 6. Note the somewhat larger difference between beginner and Intermediate groups versus intermediate and advanced groups. Considering the differences in terms of corresponding CEFR levels, the beginner group scored on average at the A2 level, the intermediate group scored on average at the lower B1 level, and the advanced group scored on average at the upper B1 level.

Similar analyses revealed the same general patterns for each of the four TOEFL Essentials test sections, with mean scores clearly distinguishing each proficiency group from the others (see Figures 7–10). Inferential tests also indicated that these patterns of difference were statistically significant for each section, as follows: reading ($F(140, 2) = 19.09$, $p = 0.000$); listening ($F(140, 2) = 28.92$, $p = 0.000$); speaking ($F(140, 2) = 22.37$, $p = 0.000$); and writing ($F(140, 2) = 18.54$, $p = 0.000$). Scores differed approximately 3 band level points between the lowest and highest proficiency groups for each section of the test, with listening section scores showing the largest overall difference (3.39 points between beginner and advanced groups).

Figure 6. Means and 95% Confidence Intervals for TOEFL Essentials Total Scores by Three Proficiency Groups

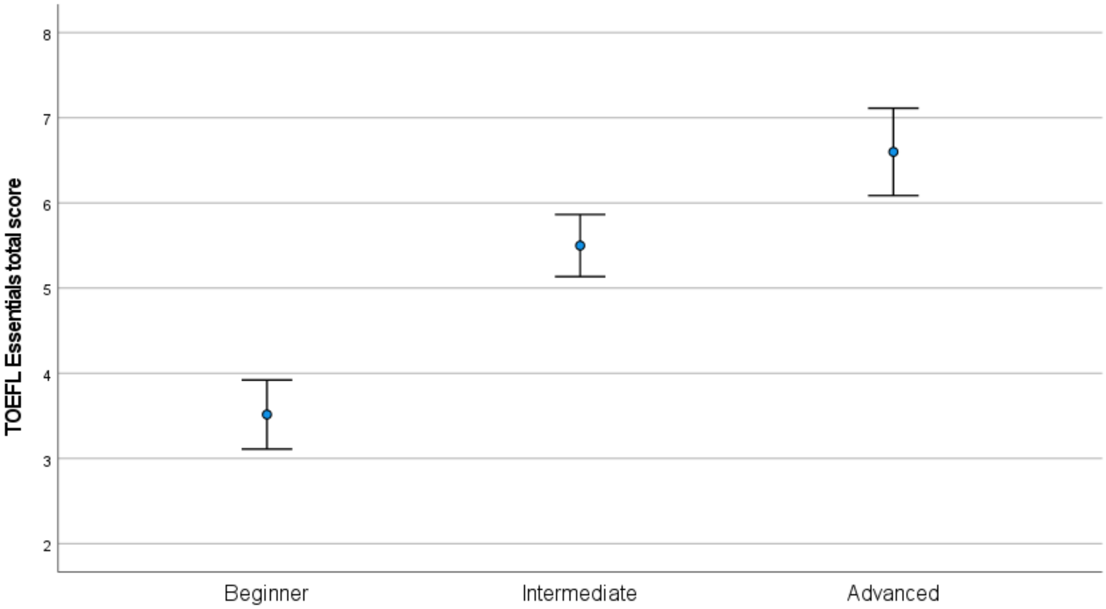


Figure 7. Means and 95% Confidence Intervals for TOEFL Essentials Reading Scores by Three Proficiency Groups

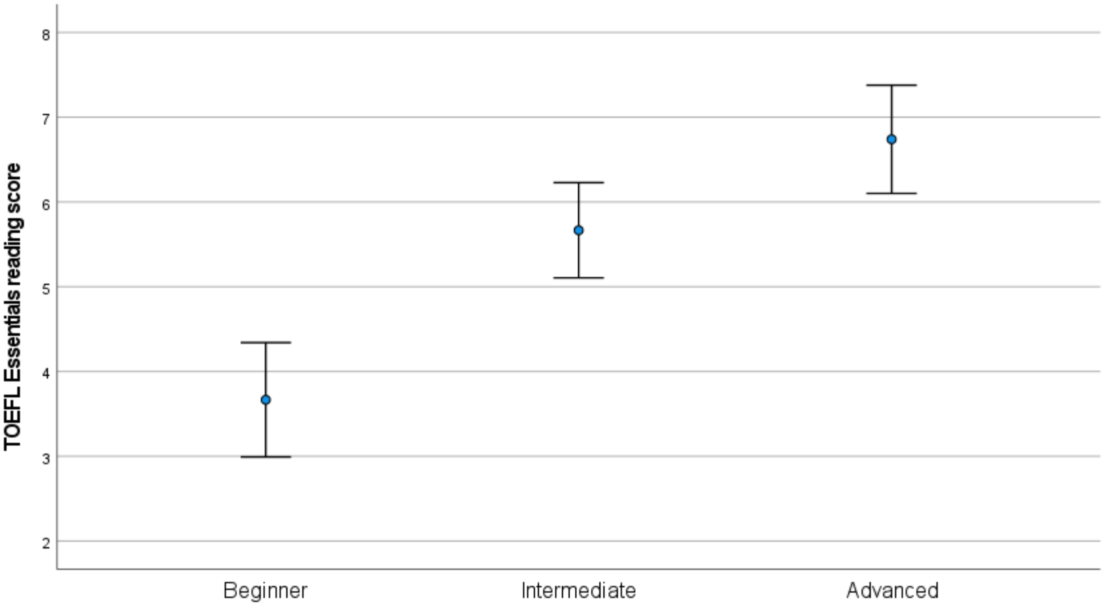


Figure 8. Means and 95% Confidence Intervals for TOEFL Essentials Listening Scores by Three Proficiency Groups

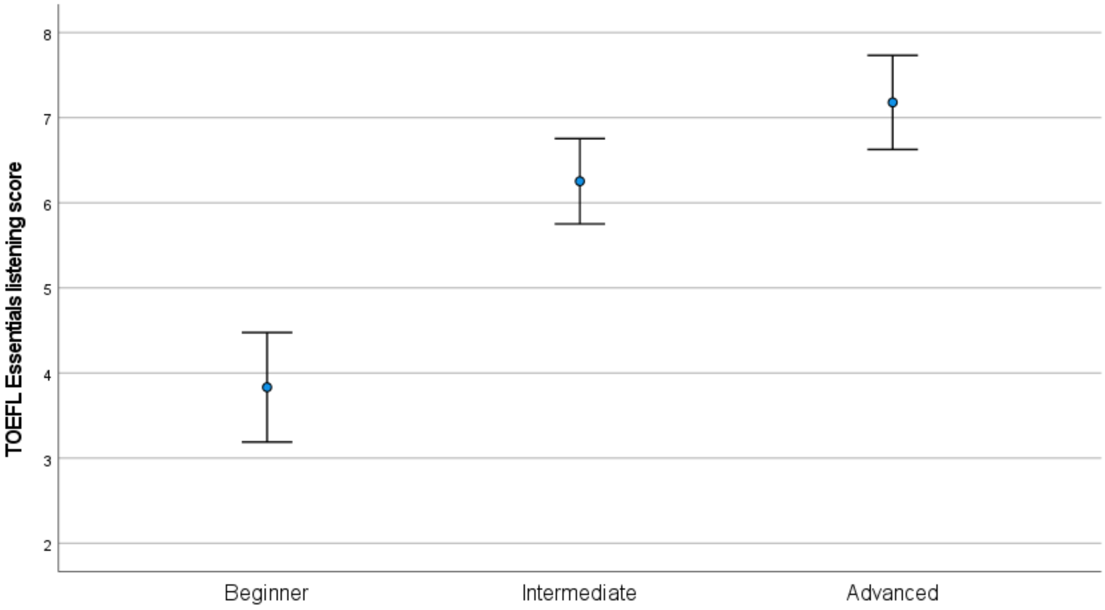


Figure 9. Means and 95% Confidence Intervals for TOEFL Essentials Speaking Scores by Three Proficiency Groups

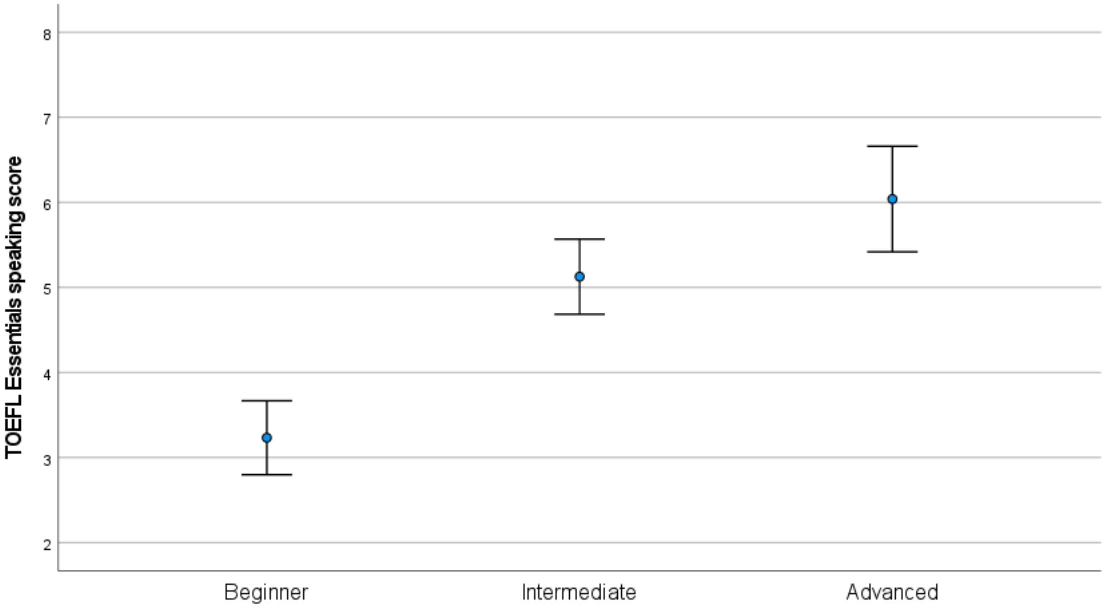
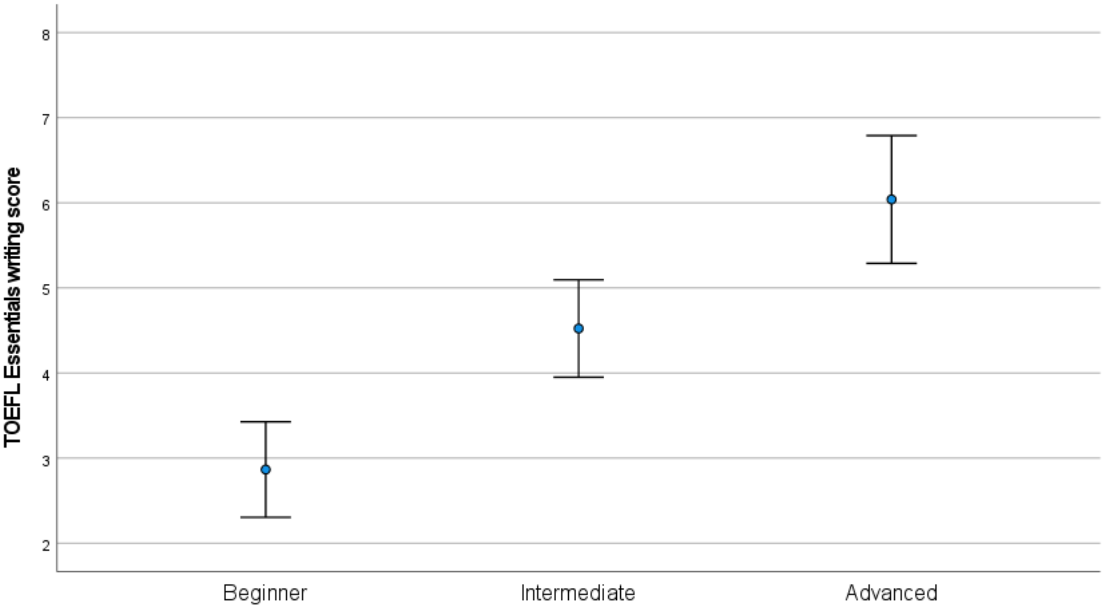


Figure 10. Means and 95% Confidence Intervals for TOEFL Essentials Writing Scores by Three Proficiency Groups



In order to gain another independent criterion measure of student proficiency, teachers at the Canadian institutions were asked to rate each learner's overall English ability according to the 12 levels of the CLB. Similar estimates were not available from the non-Canadian institutions. TOEFL Essentials total scores for the same students were then converted to their corresponding CLB levels, based on the official mapping of TOEFL Essentials test scores to CLB levels (Papageorgiou et al., 2022). Descriptive statistics for both sets of CLB level estimates are shown in Table 4 for the subsample of 113 students in Canadian institutions.

Table 4. Descriptive Statistics for Teacher and TOEFL Essentials Test Score Estimations of Canadian Language Benchmarks (CLB) Levels

Statistic	TOEFL Essentials CLB level estimations	Canadian teachers CLB level estimations
Median	6	7
Interquartile range	3	1
Min	3	2
Max	9	10

Both sources of proficiency estimations clearly spread learners out across a broad spectrum of abilities as described by the CLB. Teacher estimates ranged slightly more than did score-mapping estimates as shown by the difference in minimum and maximum scores, and the median CLB level estimation was 1 level (out of 12 total levels) higher among the teachers. A Wilcoxon signed-ranks test indicated that the TOEFL Essentials CLB level estimates were statistically significantly lower than the teacher CLB level estimates (*Z* = -5.996, *p* <.001). However, the strength of association between the two sets of CLB level estimates was also calculated using Spearman's rho, given the ordinal nature of the CLB levels, and the two distinct estimates were found to correlate strongly, r_S = .69. Similarly, the teacher estimates correlated strongly with TOEFL Essentials total band scores, r_S = .70. In sum, teachers and TOEFL Essentials test scores distinguished in similar patterns among the English learners in terms of their proficiency levels relative to each other, while the teachers estimated learners' proficiency levels slightly higher overall.

In order to examine the teacher and TOEFL Essentials score-mapping estimates of learners' CLB levels more closely, the percentages of students assigned to each level were teachers heavily utilized Level 7 in particular (40% of the students assigned there). TOEFL Essentials estimates, by contrast, distributed learners more evenly across Levels 4, 5, 6, and 7, with the highest percentage of students (27%) assigned to Level 6.

Table 5. Percentages of Teacher and Test Score Estimations of Canadian Language Benchmarks (CLB) Levels

CLB levels	TOEFL Essentials CLB level estimations	Canadian teachers CLB level estimations
2	0%	1%
3	4%	0%
4	21%	5%
5	14%	12%
6	27%	27%
7	24%	40%
8	6%	9%
9	4%	4%
10	0%	2%

Discussion

The study findings provide new evidence in support of fundamental parts of the TOEFL Essentials validity argument, and they raise a few additional questions for subsequent investigation. It is clear, first of all, that the test was quite capable of spreading learners out into a range of language abilities that were reflected by both total and section scores. The sample of English learners included those determined in advance of testing to have quite low to quite high proficiencies, and the overall performance of the learner sample on the test reflected this broad range. Importantly, the test was also quite capable of generating test scores with sufficiently high levels of reliability, indicating that the test design consistently produces scores that distinguish meaningfully among learners at different proficiency levels, even when reliability is estimated on the basis of a small sample. From the perspective of the evaluation inference in the TOEFL Essentials validity argument, then, evidence supports the claim that the test is able to consistently elicit English language performances that can be reliably scored to reflect meaningful differences in proficiency.

A more specific version of this evaluation inference, also related to the extrapolation inference, has to do with whether TOEFL Essentials test scores can distinguish effectively between learners grouped at different levels along a spectrum from low to high proficiency. The current study has provided additional evidence in support of this specific claim as well. First, mean total and section scores were found to differ consistently between three groups of learners estimated in advance of testing to have differing levels of proficiency based on information from their English language study programs. Here, it is important to note that the assignment of learners to the three different proficiency levels was complicated by the reality that (a) the proficiency ranges encapsulated by courses at each institution were somewhat uncertain and overlapping with each other; (b) because of enrollment challenges faced by the institutions, students from potentially distinct proficiency levels were assigned to the same level courses; and (c) students presented with variable English skill profiles that made it difficult to provide a single, holistic proficiency level estimate with which to assign them to the three groups. However, in spite of these complications, the average scores of each group on the TOEFL Essentials test proved robustly distinct in the order predicted. This finding provides substantial basic support for the evaluation inference that the test is effective at distinguishing between English learners at different proficiency levels. It also supports the extrapolation inference in that the comparisons were made based on real learners grouped according to in situ educational program criteria (course levels), indicating an important relationship between TOEFL Essentials test scores and real-world representations of test-takers' abilities.

Another independent source of criterion-related information regarding learner proficiency levels—investigated in relation to the extrapolation inference as well as the utilization inference—was sought through teacher estimations of English proficiency according to CLB levels for the sample of learners in Canada. These estimations were found to correlate strongly with TOEFL Essentials test scores and associated score-based CLB level estimations, providing an important additional source of evidentiary backing that test scores and teacher estimates ranked leaners in relatively similar ways. Here again, the real-world judgments of learners' proficiencies corresponded with the test scores, providing meaningful backing for extrapolation of the scores to test-taker ability in the real world.

At the same time, it was also observed that teachers tended to estimate learners' proficiency levels slightly higher than did the TOEFL Essentials test, in CLB level terms, and that teachers' judgments tended to conglomerate around a few levels on the CLB (67% of ratings at CLB 6–7), while TOEFL Essentials test scores spread learners out in a more equivalent distribution across CLB levels 4–7. It can be hypothesized that teachers may have been referring to students' course enrollment levels as one source of information for making their own estimations of learner proficiency with CLB Level 7 in particular representing an important programmatic juncture (i.e., equivalent to the transition from CEFR Level B1 to B2). It is also worth emphasizing that test takers completed the TOEFL Essentials test at the beginning of their instructional terms, and hence level estimations made by teachers in reference to the course per se may have overestimated the actual proficiencies of these learners at the beginning of the corresponding period of learning (e.g., students entering into a course at CLB Level 7 versus exiting from that course). The fact that participants' average TOEFL Essentials test scores differed noticeably less between the intermediate and advanced groups—compared with score differences between the beginning and intermediate groups—may also suggest that there was some degree of overlap in the actual proficiency levels of learners in the mid to upper course levels. Interestingly, the availability of the TOEFL Essentials score mapping to the CLB levels proved quite valuable in the current study in identifying similarities and differences in how teachers and the test assigned learners to proficiency level estimations. This affordance of the score mapping also provides some degree of support for the utilization inference, in that scores are mapped to proficiency frameworks in order to facilitate meaningful interpretation.

In sum, the current study provided new evidence from operational testing and several English language teaching and learning contexts in support of claims that the TOEFL Essentials test can distinguish effectively among learners at distinct proficiency levels. It also identified potential ambiguities in the determination of learners' proficiency levels from curricular or teacher perspectives vis-à-vis the estimation of proficiency levels on the basis of test scores, opening a new avenue for future investigations.

Limitations

Several factors limit the extent to which the current study findings support the validity of interpretations based on TOEFL Essentials test scores. While efforts were made to recruit test takers reflecting a wide range of English proficiency levels, the participant sample was relatively small, entirely voluntary, and unequally distributed across several contexts of English language study. Idiosyncratic characteristics of this small sample may not be fully reflective of the global English learner population targeted by the test, and hence patterns related to test scores in this study might not be generalizable. Participants also reflected a somewhat narrow distribution of

English proficiency levels, with the majority of learners falling in the CEFR A2, B1, and B2 levels, thereby covering only a portion of the range of proficiency levels intended to be measured by the TOEFL Essentials test. In addition, though participants were informed that their performances on the test would follow standard test-taking and proctoring protocols and result in official score reports, and they were therefore encouraged to do their best, it is possible that not all participants were motivated to engage with the test-taking experience or to demonstrate their full English proficiency across all tasks. Perhaps the most critical limitation to the study findings has to do with uncertainties regarding the accuracy of selected criterion variables, in this case both the course level and teacher judgment variables. As reported by the participating institutions, course level was at best a rough proxy for English proficiency, both because of the uneven skill profiles of the participants and because learners with distinct proficiencies were placed into the same course levels due to enrollment policies. Similar concerns with teacher judgments reflect the challenge they faced in accurately estimating individual learners' English proficiency levels, across the four language skills and holistically, and in reference to potentially less familiar frameworks like the CEFR or CLB. These limitations should be kept in mind when considering the implications of the current study.

Conclusion

Limitations notwithstanding, evidence collected in the current study clearly supports the inference that both section and total scores on the TOEFL Essentials test are effective at distinguishing among learners at different English proficiency levels. Performance data revealed broad distributions of section and total test scores, reflecting the test's capacity to capture the wide range of English abilities among the groups of students recruited from distinct course levels and estimated to vary in their proficiency by language teachers. Test scores also exhibited high internal consistency reliability estimates for this relatively small and idiosyncratic sample of learners, highlighting the robust psychometric foundations of the test design. Moreover, mean scores for beginner, intermediate, and advanced proficiency groupings (based on a priori course level differences) differed in clear and statistically significant degrees in the order predicted and on the four skill sections as well as the total test. Lastly, for the subgroup of students in Canadian institutions, teacher ratings of the overall CLB levels of students correlated strongly with students' total test scores, providing an independent criterion in support of the test's capacity to distinguish English proficiency levels.

Interestingly, when estimations of CLB levels were made by the teachers in comparison with the same estimations based on score-mapping of the TOEFL Essentials test scores with the CLB, a generalized pattern of slight discrepancy was identified. That is, teachers' estimated learner English proficiency slightly higher on the whole than did the test scores as mapped to the CLB framework. This observation bears future consideration, both as a focus of investigation for the test, in order to verify claims regarding score-mapping to the CLB, but also as a potential point of evaluation for the teachers and programs, in order to evaluate the alignment of teachers' perceptions of proficiency in relation to level descriptions in the CLB (to which the courses and curricula are ostensibly aligned). Another interesting pattern had to do with the generally uneven language skill profiles exhibited by the sample of students, with overall lower writing and speaking abilities in comparison with listening and reading abilities. While this pattern may reflect a known tendency among learners who come from instructional contexts that emphasize the teaching of receptive skills, it is nevertheless of interest for ongoing monitoring, both from a test validity perspective (is the test adequately estimating abilities on each of the four language skills?) as well as from an English language program perspective (are students showing up with uneven skill profiles, and how does instruction respond to that reality?).

Ultimately, the evidence gathered in the current study provides some initial backing, gathered on the operational test from real test takers in an actual English language learning context, in support of the inferential claim that the TOEFL Essentials test scores are able to effectively and reliably distinguish among English learners at distinct proficiency levels (evaluation inference) and that the test scores are related to real-world indicators of language proficiency (extrapolation and utilization inferences). Additional research is, of course, called for to look beyond these claims at additional inferences in the overall validity argument for the test.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill College.

Centre for Canadian Language Benchmarks. (2012). *Canadian Language Benchmarks: English as a second language for adults*. https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19–27. https://doi.org/10.1177/0265532211417211

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

Chapelle, C. A., & Lee, H.-w. (2021). Conceptions of validity. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 17–31). Routledge. https://doi.org/10.4324/9781003220756-3

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press. https://doi.org/10.1017/9781108669849

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

Davis, L., Garcia Gomez, P., Li, S., & Manna, V. F. (2023). Mapping TOEFL Essentials speaking and writing scores to the CEFR Levels. In S. Papageorgiou & V. F. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation* (pp. 120–140). John Benjamins. https://doi.org/10.1075/illa.1.07dav

Davis, L., Norris, J., Papageorgiou, S., & Sasayama, S. (2023). Balancing construct coverage and efficiency: Test design, security, and validation considerations for a remote-proctored online language test. In K. Sadeghi & D. Douglas (Eds.), *Fundamental considerations in technology mediated language assessment* (pp. 49–63). Routledge. https://doi.org/10.4324/9781003292395-5

Fleckenstein, J., Leucht, M., & Köller, O. (2018). Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly, 15*(1), 90–101. https://doi.org/10.1080/15434303.2017.1421956

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Norris, J. M. (2008). *Validity evaluation in language assessment*. Peter Lang. https://doi.org/10.3726/978-3-653-01171-5

North, B., & Piccardo, E. (2018). *Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of References (CEFR)*. Centre for Canadian Language Benchmarks.

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL Junior standard scores for ESL placement decisions in secondary education. *Language Testing, 31*(2)*,* 223–239. https://doi.org/10.1177/0265532213499750

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL Essentials test 2021* (Research Memorandum No. RM-21-03). ETS. https://www.ets.org/Media/Research/pdf/RM-21-03.pdf

Papageorgiou, S., Davis, L., Ohta, R., & Garcia Gomez, P. (2022). *Mapping TOEFL Essentials test scores to the Canadian Language Benchmarks* (Research Report No. RR-22-16). ETS. https://doi.org/10.1002/ets2.12357

Notes

[1] Note that reliability estimates were not calculated for TOEFL Essentials section scores, as item-level response data were not available for their calculation; however, given that each section score contributed to the estimation of an overall quite high test score reliability, it can be inferred that the four sections produced scores with sufficient levels of consistency.