

Test-Taker Perspectives on English Proficiency Test Predictive Validity

Samuel J. Miller

A Master's Thesis

Submitted to the Department of Curriculum and Instruction, College of Education and Human
Development, University of Minnesota, Twin Cities

Advisor: Samuel S. David

May 2026

Acknowledgements

The author wishes to thank all those who helped support this research, which was started in September 2025 and completed in May 2026. Special thanks go to Samuel S. David, the author's academic advisor, for his direction throughout the project, including the initial topic, ideas, the time, the IRB process, the survey and interview format, editing of the thesis and questioning of the content for revision, and final review of the paper. In addition, the author thanks the University of Minnesota Department of Curriculum and Instruction Graduate Program Coordinator for his support in finding students in the department and connections to campus departments and organizations to assist in the dissemination of the test-taker survey.

Thank you also to each of the test-takers for completing the survey and being willing to share their perceptions of the proficiency tests, as well as the two interviewees for enthusiastically sharing their testing experiences. All these contributions made this project and paper possible.

Finally, the author wishes to thank the Lord Jesus Christ for giving him the strength and endurance to complete this project and do it to the best of his ability.

Abstract

This study evaluates the predictive validity of English proficiency tests—specifically the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) and the Duolingo English Test (DET)—in determining the academic success of international students. While traditional assessments like TOEFL and IELTS are well-established, the rapid rise of digital-first alternatives like the DET necessitates further investigation into their effectiveness and impact on test-taker confidence.

Through a mixed-methods approach, this study reviews existing quantitative research on test score correlations with Grade Point Average (GPA) and introduces original qualitative data from an anonymous survey and interviews with students at the University of Minnesota. The findings indicate that while higher proficiency scores generally relate to increased language confidence and belief in test accuracy, individual experiences vary significantly. Qualitative case studies reveal that high test scores do not always equate to immediate classroom readiness, highlighting a potential gap between test performance and real-world academic demands. Ultimately, the research suggests that while the DET is a legitimate alternative of predictive validity to traditional tests, stakeholders should consider test-taker perceptions and holistic preparation factors rather than relying solely on proficiency scores for admission decisions.

Table of Contents

Acknowledgements	2
Abstract	3
Table of Contents	4
List of Tables	6
Test-Taker Perspectives on Proficiency Test Predictive Validity	7
Literature Review	9
History of Modern Language Testing	9
TOEFL History	11
TOEFL Structure	13
IELTS History	15
IELTS Structure	17
DET History	18
DET Structure	19
Comparing TOEFL, IELTS and DET	20
TOEFL, IELTS, and DET Growth Comparison	21
Purpose of Proficiency Tests	26
Validity in Proficiency Testing	29
Predictive Validity of TOEFL	30
Predictive Validity of IELTS	32
Predictive Validity of DET	34
Methodology	37
Study Research Questions	38
Research Design	38
Study Procedure	39
Data Preparation	41
Results	43
Confidence and Test Scores	43
Accuracy and Test Scores	49
Discussion & Conclusion	61
Limitations	68
References	71
Appendix A	84
English Proficiency Test Survey	84
Appendix B	86
Interview Questions	86
Appendix C	87
Data Point Combinations for Analysis	87

List of Tables

Table 1: Minimum English Proficiency Score Requirements for Admission	7
Table 2: Comparison of ELTS and IELTS Exam Modules	16
Table 3: TOEFL, IELTS, and DET Feature Comparison	20
Table 4: TOEFL, IELTS, and DET Usage Growth	22
Table 5: Count of U of M English Proficiency Submissions by Test Type and Academic Year, 2020/21-2024/25	25
Table 6: Test-Taker Majors and Categorizations	41
Table 7: CEFR Scores Overall and Each Skill	43
Table 8: Confidence Level for Each Skill	44
Table 9: CEFR Scores and Confidence Level	45
Table 10: CEFR Scores By Test	46
Table 11: Confidence Level for Each Skill by Test	47
Table 12: CEFR Scores and Confidence Level by Test	48
Table 13: Accuracy Feeling by CEFR	50
Table 14: Accuracy Feeling by CEFR Subscore	50
Table 15: Accuracy Feeling by Test	52
Table 16: CEFR Scores and Accuracy Feeling by Test	52
Table 17: CEFR Subscores and Accuracy Feeling by Test	53
Table 18: Accuracy Feeling Totals by Subscore and Test	54
Table 19: Test Accuracy Feeling vs. Preparatory Ability	57
Table 20: CEFR vs. Preparatory Ability	58
Table 21: TT1 vs. TT2 Test Experience	67

Test-Taker Perspectives on Proficiency Test Predictive Validity

Students wanting to pursue an education at an English-medium university navigate many challenges in the study abroad preparation process. One of the significant obstacles is providing proof of their linguistic readiness for academic study in English.

International students applying to the University of Minnesota, for example, will see the following English proficiency test score minimum requirements for admission:

Table 1:

Minimum English Proficiency Score Requirements for Admission

English Proficiency Test Name	Test Score	Writing Score
TOEFL iBT	79	21
IELTS Academic	6.5	6.5
Pearson Test of Academic English (PTE Academic)	59	59
Cambridge Assessment English C1 Advanced (CAE)	180	180
Duolingo	110	110

Note: ([University of Minnesota, n.d.](#)) Scores listed above represent minimums for the university, but individual graduate programs may set higher standards. Students with scores below the listed minimum may obtain conditional admission, requiring English class attendance before enrollment.

This paper traces the history of English proficiency tests, providing an in-depth comparative analysis of the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) and the Duolingo English Test (DET). While all three tests are included in Table 1, they represent fundamentally different philosophies of test design. IELTS and TOEFL, developed during the mid-to-late 20th century, are established benchmarks originally designed for paper-and-pencil delivery in proctored testing centers; both have since achieved widespread global acceptance ([International English Language Testing System, 2017](#); [Chapelle et al., 2008](#)). In contrast, the DET, launched in 2016, is a modern, digital-first

proficiency test that can be taken remotely from any location. Despite its relatively recent launch, its global adoption has accelerated rapidly ([Issuu, 2022](#)).

Each of these tests, accepted by universities worldwide, is globally recognized benchmarks of English language proficiency ([British Council, 2024](#); [Naismith et al., n.d.](#); [Chapelle et al., 2008](#)). This paper not only provides a head-to-head comparison of design and features, but also evaluates the research about their effectiveness. This study examines whether a high score on TOEFL, IELTS or DET is a reliable predictor of a student's future academic success. By reviewing previous research, the relative effectiveness of all models can be assessed.

Developing a comprehensive understanding of the practical implications of English proficiency scores is essential for all stakeholders within the educational ecosystem. For test developers, the critical task lies in designing assessments that accurately mirror the complex communicative skills required in university settings, thereby ensuring that scores truly represent a student's linguistic capacity ([McNamara, 2011](#)). Concurrently, university administrators must recognize that a standardized proficiency score is not a holistic indicator of a student's overall readiness for higher education ([Ihlenfeldt & Rios, 2022](#)). Finally, because students and their families invest substantial time, emotional energy, and financial resources into preparing for and sitting for these assessments, they must understand the limitations inherent to these metrics; achieving a target benchmark does not inherently guarantee academic success ([Andrade, 2006](#); [Dawadi 2020](#)).

Beyond describing some of the quantitative research of previous studies, this paper also introduces qualitative research conducted at the University of Minnesota through anonymous surveys and student interviews. Quantitative research on English proficiency tests gives us a

macro-level perspective of the trends in the data, while qualitative research provides a more nuanced understanding of underlying factors of those trends. This study's intent is to capture test-takers' own perceptions about how proficiency tests shape academic confidence and actual experience.

Extensive research exists on the validity of TOEFL and IELTS tests, due to their longevity and popularity worldwide. The DET test, on the other hand, while having some internal reports and third-party research available, remains comparatively limited, mainly because the test has been in circulation for a relatively short period of time. This study addresses this scholarly gap by providing qualitative data, analyzing how test-taker experiences and survey results reflect DET's practical utility.

This paper first examines the history of English proficiency testing, highlighting how major developments in the understanding of language have shaped the development of tests. It then provides a comparative analysis of TOEFL, IELTS and DET, including the history of development, test structure and format, and a summary of test validity research findings from previous research. Subsequently, the study presents original research with its methodology, results, and discussion. The paper concludes by connecting previous research with current data to discuss the implications for international education and admissions policy.

Literature Review

History of Modern Language Testing

The pedigree of language assessment spans millennia. [Spolsky \(2008\)](#) identifies the Chinese imperial examinations as the earliest instance of centralized testing, holding Chinese imperial examinations to find the best candidates for government office. Initiated during the Sui

dynasty (581-618 CE) and reaching full maturity during the Qing dynasty, these exams set a precedent for high-stakes assessment ([Cohen & Teiser, 2025](#)). In the West, the shift toward English proficiency began in the 15th century, catalyzed by an ordinance from Henry V that transitioned royal correspondence from French to English ([Giri, 2010](#)).

The 20th century marked the professionalization of the field. [Mao \(2022\)](#) highlights Spolsky's influential tripartite categorization of modern language testing: the *Pre-scientific* period (prior to the 1950s), the *Psychometric-structuralist* period (1950s-1960s), and the *Psycholinguistic-sociolinguistic* period (1960s-present). As linguistic theory advanced, testing methodologies shifted to reflect new understandings of how language is acquired and utilized.

The *Pre-scientific* period focused on grammatical, lexical, and phonetic accuracy, mirroring the 'grammar-translation' pedagogy of the time ([Farhady, 2018](#)). Tests were created and analyzed by individual teachers and subject to their interpretation of the results, lacking any standardized theoretical basis. This changed with the advent of the *Psychometric-structuralist* movement, which prioritized reliability and empirical validity. Rooted in the 'audiolingual method', [Lado \(1961\)](#) posited that language ability was the sum of its discrete parts ([Farhady, 2018](#)). This gave rise to *discrete-point testing*—typically via multiple choice formats—which provided the standardized reliability necessary for the birth of industrial-scale exams like the TOEFL and IELTS. ([Spolsky, 2008](#); [Mao, 2022](#))

Reaction against the fragmentation of language led to the *Psycholinguistic-sociolinguistic* shift. [Oller \(1979\)](#) challenged the discrete point model with his 'Unitary Competence Hypothesis' (UCH), arguing that language is an integrated system rather than a collection of isolated skills ([Giri, 2010](#)). Consequently, testing became more integrative, requiring candidates to synthesize multiple linguistic abilities simultaneously. By the 1980s, this

evolved into *Communicative Language Ability (CLA)* which incorporated social and strategic competencies ([Canale & Swain, 1980](#); [Farhady, 2018](#)).

Under the CLA framework, testing focused on four pillars: *real-life tasks, face validity, authenticity, and performance* ([Fulcher, 2000](#), p. 484). Essentially, an exam must present relevant, clear prompts that mirror practical usage and evaluate actual linguistic performance. While this ‘re-humanized’ the assessment process, it introduced a new tension: balancing the ‘authenticity’ of real-world tasks with rigorous ‘reliability’ previously offered by multiple-choice formats ([Fulcher, 2000](#)).

[Morrow \(1979\)](#) famously metaphorized this evolution as a journey from the “Garden of Eden” through the “Vale of Tears” to the “Promised Land”—a progression from simplistic comfort to a complex, hard-won haven of understanding. Yet, the “Promised Land” remains an evolving target. As [Fulcher \(2000\)](#) notes, language tests are not perfect mirrors of reality; they are sophisticated instruments constructed upon our best current theories of language and learning.

TOEFL History

Test of English as a Foreign Language (TOEFL) was the earliest of the three major English proficiency assessments to be launched. Developed in the early 1960s, the TOEFL emerged from a cooperative effort of more than thirty public and private institutions concerned with evaluating the linguistic readiness of non-native speakers applying to English-medium schools. Initially financed by grants from the Ford and Danforth Foundations, the program shifted in 1965 when the College Board and Educational Testing Service (ETS) assumed joint administrative responsibility for the assessment.

Following its initial launch in 1964, the TOEFL underwent three distinct developmental stages over its first forty years of operational use. Throughout this evolution, not only was its delivery infrastructure updated, but the underlying construct and content of the exam were systematically re-engineered to align with innovations in language proficiency theory. The inaugural stage (1964–1979) operated under a structuralist testing paradigm, focusing almost exclusively on discrete language components measured through multiple-choice matrices. Although developers recognized that productive skills like speaking and writing should ideally be assessed, the testing industry lacked the large-scale logistical frameworks required to achieve this at the time. Delivered strictly via paper-and-pencil formats during this period, annual test-taker volume was estimated between 40,000 and 150,000 candidates ([Spolsky, 1995](#)).

The second stage of development (1979–2005) marked a transition into the communicative testing movement, during which ETS introduced direct measures of productive competence to complement the existing multiple-choice reading and listening sections. Coinciding with a massive explosion in East Asian and European student mobility, annual test-taker volume expanded dramatically to between 400,000 and 800,000 candidates ([Stoynoff, 2009](#)). These new productive assessments were deployed as independent auxiliary modules: the Test of Spoken English (TSE) in 1980 and the Test of Written English (TWE) 1986, which were taken by a smaller subset of candidates to fulfill specific institutional requirements. It was also during this stage, in 1998, that the first localized computer-based version of the exam (TOEFL CBT), was introduced.

Finally, the contemporary era of the assessment began in 2005 with the launch of the Internet-Based Test (TOEFL iBT). Beyond utilizing a new remote delivery architecture, the iBT framework was explicitly mapped to the CLA model, shifting the test's core construct toward the functional integration of receptive and productive skills to measure authentic *communicative*

competence ([Bachman, 1990](#); [Education Testing Service, 2023](#)). The task-centric design was guided by the operational philosophy that test-takers must be evaluated using integrated scenarios and contextualized tasks that directly mirror academic life in higher education ([Education Testing Service, 2023](#)). Driven by this authentic architecture and widespread institutional adoption, annual registration volumes crossed the 1 million mark during this period ([Stoynoff, 2009](#)).

TOEFL Structure

The TOEFL iBT exam, introduced by ETS in 2005, represents a linear, non-adaptive assessment designed to measure communicative competence across four primary linguistic domains: reading, listening, speaking and writing. A distinct structural divergence exists between the TOEFL iBT and competing instruments such as the IELTS. Whereas the IELTS isolates oral proficiency through an independent, face-to-face interview with an examiner, the TOEFL iBT operationalizes the speaking construct via a computer-delivered interface using a headset, embedding the section directly within the collective administration sequence.

Historically, the structural architecture of the TOEFL iBT was governed by the construct of *sustained academic stamina* to validate a candidate's capacity to withstand the cognitive demands of higher education environments ([Chapelle et al., 2008](#)). Consequently, for nearly fifteen years following its inception, the instrument was characterized by a protracted administration time of 3.5 to 4 hours, bifurcated only by a mandatory ten-minute intermission.

Under this initial framework, the exam components were delineated as follows:

- Reading: Composed of three to four academic passages of approximately 700 words each, requiring candidates to respond to 12 to 14 items per passage within a

54-to-72-minute window to evaluate literal comprehension, inferential reasoning, and contextual vocabulary.

- Listening: Appraised main-idea comprehension, pragmatic understanding, and information synthesis over 41-57 minutes through three to four conversational tracks (5 items each) and four to six academic lectures (six items each).
- Speaking: Spanned 20 minutes and necessitated the execution of six distinct tasks, categorized into two independent tasks evaluating opinions on familiar campus or personal dilemmas, and four integrated tasks requiring the verbal synthesis of prior reading and listening stimuli.
- Writing: Allocated 50 minutes for the production of two compositions: an integrated synthesis essay (150-225 words) deriving from multi-modal inputs, and a traditional independent argumentative essay (minimum 300 words).

The temporal fluctuations observed in the historical Reading and Listening sections were an artifact of experimental research design. ETS systematically embedded unscored variable items—frequently designated as ‘dummy questions’--to pilot future test specifications. Because these items were indistinguishable from active test content, candidates were required to maintain peak cognitive performance across all tasks to safeguard their composite scores ([Educational Testing Service, 2020](#)).

While the underlying construct remained relatively static for a decade and a half, the testing instrument has undergone iterative revisions since 2019 to optimize efficiency and reduce candidate fatigue. In August 2019, ETS condensed the examination by approximately 30 minutes. This efficiency was achieved by curtailing the volume of reading items and eliminating specific tasks within the listening and speaking domains, though the macro-level construct and psychometric scoring mechanics remained unaltered ([Sudarsono, 2021](#)).

The most definitive paradigm shift occurred In July 2023, when ETS significantly compressed the total test duration to under two hours. This reconfiguration systematically dismantled several long-standing features of the exam; all unscored experimental items were eliminated, the mandatory ten-minute intermission was excised, and the reading domain was restricted to two academic passages featuring 10 items each. Crucially, the independent writing task was obsolete, replaced instead by a highly accelerated, 10-minute 'Writing for an Academic Discussion' task ([Educational Testing Service, 2023](#)). This ongoing structural compression suggests a theoretical re-evaluation of how academic readiness is operationalized, pivoting away from historical measures of sustained cognitive endurance toward more localized, high-efficiency language tasks.

TOEFL iBT performance outcomes are reported on a continuous integer scale in one-point increments, yielding a maximum aggregate score of 120 points. Each of the four discrete linguistic sub-scales—reading, listening, speaking, and writing—is weighted equally, accounting for a maximum allocation of 30 points per section ([Educational Testing Service, 2020](#)). Institutional matriculation thresholds vary; for instance, as delineated in Table 1, the University of Minnesota mandates a minimum composite score of 79 for unconditional undergraduate admission ([University of Minnesota, n.d.](#)).

IELTS History

While the IELTS officially debuted in 1989, its conceptual roots extend back to the 1960s. In 1963, a collaboration between the British Council and the University of Birmingham produced the English Proficiency Test Battery (EPTB). Launched in 1965, the EPTB focused exclusively on receptive skills—reading and listening—which soon proved insufficient as the field shifted toward communicative language assessment.

To address these limitations, Cambridge English Language Assessment (UCLES) and the British Council developed the English Language Testing Service (ELTS) in 1980. Though more comprehensive, the ELTS was hindered by its modular complexity, featuring six distinct academic tracks ranging from Life Sciences to Technology. This specialized approach proved logistically taxing and redundant; by 1985, annual participation stalled at just 10,000 test-takers ([Manhattan Review, n.d.](#)).

A major institutional redesign followed, incorporating the International Development Program of Australian Universities and Colleges (IDP). This partnership signaled a transition from a UK-centric model to a global one, rebranded as 'IELTS.' The 1989 relaunch replaced the six specialized modules with the standardized four-domain framework (reading, writing, listening, and speaking) used today. This streamlined design catalyzed a surge in popularity: by 1995, annual participation reached 43,000 across 210 global testing centers ([Manhattan Review, n.d.](#); [International English Language Testing System, 2017](#)).

Table 2:

Comparison of ELTS and IELTS Exam Modules

ELTS Modules	IELTS Modules
Life Science	Reading
Social Studies	Writing
Physical Science	Listening
Technology	Speaking
Medicine	
General Academic Assessment	

IELTS Structure

Since its inception, the IELTS has maintained a consistent structure centered on four core modules: Listening, Reading, Writing, and Speaking. While the exam offers both General Training and Academic versions, this paper focuses exclusively on the *IELTS Academic*, the standard for university admissions.

The total examination time is approximately two hours and forty-five minutes. Notably, the Speaking Module is conducted as a separate face-to-face interview, which may be scheduled up to one week apart from written components.

- **Listening:** This 30-minute module comprises four sections of increasing complexity, ranging from social interactions to academic lectures. Candidates are provided an additional ten minutes to transfer responses from the test booklet to the formal answer sheet.
- **Reading:** Over 60 minutes, test-takers analyze three non-specialist texts totaling 2150 to 3250 words. Assessment tasks vary from multiple choice to diagram labeling; notably, accuracy in spelling and grammar is a prerequisite for credit.
- **Writing:** This 60-minute module requires two distinct tasks: a data description exercise involving a visual prompt (Task 1) and a formal argumentative essay (Task 2).
- **Speaking:** A 10-to-15-minute proctored interview consisting of a personal introduction, a sustained monologue on a designated prompt, and a conceptual discussion with the examiner.

Results are reported via a 'Band Scale', ranging from 1.0 to 9.0 in half-band increments. Both individual module subscores and a cumulative 'Overall Band Score' are provided

([Manhattan Review, n.d.](#); [International English Language Testing System, 2017](#)). As illustrated in Table 1, the University of Minnesota mandates a minimum overall score of 6.5 for unconditional international admission ([University of Minnesota, n.d.](#)).

DET History

The Duolingo English Test (DET) was conceived by software developer Luis von Ahn and computer scientist Severin Hacker, also the founders of the Duolingo language learning platform. Drawing from their personal experiences as international students, von Ahn and Hacker sought to democratize language assessment by removing the financial and logistical barriers associated with traditional exams. The cost of the DET is much less than IELTS or TOEFL and can be taken at home, eliminating the need to travel and schedule a time to sit at a test center ([Ramanathan, 2025](#); [Prodigy Finance, 2023](#)). For a head-to-head comparison with TOEFL and IELTS, see Table 3 below.

In 2014, the company introduced ‘Test Center’, a digital language certification program that provided users with a proficiency rating on a scale of one to ten. While initially a peripheral project, the high demand for accessible certification—coupled with the realization that the testing industry was a rigid monopoly—prompted the founders to pivot toward a direct competitor for established exams like the TOEFL and IELTS ([Lardinois, 2014](#)).

The DET officially launched in 2016 as the world’s first digital-first Secure English Language Test (SELT). Designed for maximum accessibility, the exam requires only a webcam-equipped computer, a stable internet connection, and is relatively inexpensive as compared to IELTS or TOEFL ([Prodigy Finance, 2023](#); [Svensson, 2025](#)). The test’s adoption catalyzed in 2020 when the COVID-19 pandemic forced the closure of physical testing centers. As an on-demand, home-based alternative, the DET filled a critical vacuum, gaining acceptance

from 2,000 new institutions in that year alone ([Issuu, 2022](#)). By 2025, the exam's institutional footprint expanded to over 6,000 organizations, including all Ivy League universities and nearly the entire top tier of global higher education ([Svensson, 2025](#)).

DET Structure

Although TOEFL, IELTS and DET aim to evaluate the same fundamental linguistic competencies, their underlying architectures are vastly different. The IELTS, established in 1989, was originally a linear paper-based exam; in contrast, the DET debuted in 2016 as a digital-first assessment, leveraging nearly three decades of advancements in educational technology.

The DET is a streamlined, one-hour examination divided into three distinct phases: a five-minute onboarding session, a 45-minute adaptive core, and a ten-minute writing and speaking sample. During the onboarding phase, the system verifies the candidate's identity and ensures technical compatibility. Security is maintained through a hybrid proctoring model, utilizing human reviewers assisted by AI-driven monitoring to enforce strict integrity protocols in a remote environment.

The core of the DET utilizes *Computer Adaptive Testing (CAT)*, where AI algorithms adjust the difficulty of the subsequent items based on the test-taker's previous performance. Notably, the DET is the inaugural high-stakes assessment to utilize generative AI for automated item generation, ensuring that an individual would need to take the exam over 1,000 times to encounter a repeated question ([Burstein et al., n.d.](#); [Duolingo, 2024](#)). The exam features 15 distinct task types, many of which require multimodal integration (e.g., 'Listen and Respond' or 'Read and Complete').

The final phase produces a speaking and writing portfolio for university admissions officers. Candidates must provide sustained responses—lasting up to five minutes—to specific prompts. Scoring ranges from 10 to 160 in five-point increments, aligned with the Common European Framework of Reference (CEFR). Beyond the overall holistic score, the DET provides eight subscores. Four subscores represent traditional domains, while the remaining four – Production, Literacy, Comprehension, and Conversation—measure integrated skills, reflecting the multifaceted way language is used in academic practice ([Duolingo, 2024](#)). As indicated in Table 1, the University of Minnesota requires a minimum overall score of 110 for unconditional admittance ([University of Minnesota. n.d.](#)).

Comparing TOEFL, IELTS and DET

Table 3 compares a number of TOEFL, IELTS and DET key characteristics side-by-side:

Table 3:

TOEFL, IELTS, and DET Feature Comparison

Criteria	TOEFL	IELTS	DET
Year of launch	1964	1989	2016
Test acceptance	13,000+ institutions	12,000+ institutions	6,000+ institutions
Test environment	Test center	Test center, limited online	All online
Test price	\$173-475 (country dependent)	\$215-345 (country dependent)	\$70
Test time	3hr 30m	2hr 45m	1hr
Test sections	4 - Speaking, Writing, Reading, Listening	4 - Speaking, Writing, Reading, Listening	3 - Introduction, Adaptive Test, Writing/Speaking Sample
Test questions	24-39 Listening, 30-40 Reading, 2 Writing, 4 Speaking	40 Listening, 40 Reading, 2 Writing, 3 Speaking	19 question types, may be repeated adaptively

Scoring scale	1-120 (1-point increments)	1-9 (0.5 band increments)	10-160 (5-point increments)
Score report	Overall score + 4 subscores	Overall score + 4 subscores	Overall score + 8 subscores
Results	72 hours	13 days (test center) 3-5 days (online)	48 hours
Send scores free	4 schools	5 schools	unlimited

Note: TOEFL iBT comparison employs the August 2019 edition for Test questions and Scoring scale.

The above table provides a foundational comparison of the three proficiency tests, utilizing data from various contemporary sources for the DET ([Cooper, 2023](#); [Duolingo, 2024](#); [Svensson, 2025](#)), IELTS ([Amazing Talker, 2026](#); [Cooper, 2023](#); [International English Language Testing System, 2017](#); [IELTS, 2024](#)) and TOEFL ([Educational Testing Service, n.d.-a](#); [Educational Testing Service, n.d.-b](#); [Educational Testing Service, n.d.-c](#); [Educational Testing Service, 2020](#); [Study.com, 2026](#)). These sources represent the typical information a student would encounter when researching test options.

Initial comparisons often suggest a clear preference. [Cooper \(2023\)](#) addresses the dilemma directly, stating, “The short answer is: Only take IELTS if you have to.” Similarly, [Lumetest \(2025\)](#) concludes that if an institution accepts both, “the Duolingo English Test (DET) is a cheaper and more convenient option.” While such advice appears practical, it overlooks the most critical factor for any candidate: test validity. Before addressing whether these assessments truly measure what they claim to, further structural comparisons are necessary.

TOEFL, IELTS, and DET Growth Comparison

Examining adoption rates over time provides critical insight into a standardized test’s market impact and scalability. Launched in 2016, the DET has been active for nine years at the

time of this writing. However, DET, as well as TOEFL and IELTS, all do not publish comprehensive annual usage statistics. Consequently, operational scales must be deduced from data disclosed during specific milestone years; for example, the most recent DET baseline relies on calculated estimates from 2024. To provide an equivalent perspective on their respective trajectories, Table 4 pairs current DET estimates against TOEFL and IELTS usage data from their own eighth year of operation.

Table 4:

TOEFL, IELTS, and DET Usage Growth

Test Name	Start Year	Year Eight / Usage
TOEFL	1964	1972 / 57,500
IELTS	1989	1997 / 60,000
DET	2016	2024 / 700,000

Historical candidature trajectories demonstrate the vast expansion of international English language assessment programs over the past fifty years. Longitudinal volume analysis of archival ETS data reveals that during the late 1970s, the TOEFL program averaged approximately 78,000 to 80,000 registrants annually ([Educational Testing Service, 1982](#)). Based on these baselines, language assessment historians estimate that the operational volume in 1972 was comparatively lower, situating the global examinee population between 50,000 and 65,000 registrants ([Educational Testing Service, 1982](#); [Hale et al., 1984](#)).

By comparison, the market introduction of competing instruments established distinct growth trajectories. Following the initial launch of the IELTS in 1989, candidate

volume expanded at an annualized rate of approximately 15%, culminating in an active testing population exceeding 43,000 by 1995 ([Cambridge University Press & Assessment, 2017](#)). This rapid adaptation curve persisted through the end of the decade; [Boddy \(2001\)](#) observed that “over 100,000 took [the IELTS]” during the 1999 calendar year (p. 2). Adjusting for this growth curve through mathematical interpolation between these empirical milestones yields a calculated estimate of approximately 60,000 candidates for the 1997 testing cycle.

Conversely, contemporary digital-first assessment models reflect an entirely different operational paradigm and structural scale. Utilizing corporate monetization frameworks and public revenue streams, an industry analysis by [Arno \(2026\)](#) estimates that approximately 700,000 candidates sat for the DET in 2024 alone. This steep comparative growth indicates a significant macro-level shift in the global language testing market, characterized by an exponential acceleration in candidate volume driven by digitized, decoupled remote proctoring systems.

While these disparate adoption figures suggest that the DET achieved a more accelerated initial growth trajectory, this divergence must be evaluated against several critical contextual and systematic discrepancies. Most fundamentally, these instruments were conceptualized and deployed across vastly different technological milieus. At the inception of the TOEFL in 1964, and the IELTS in 1989, high-stakes language assessment infrastructure was exclusively paper-based, legally and logistically binding candidates to physical administration centers.

The transition to digital delivery models required decades of structural adaptation for legacy providers. ETS did not implement a computer-based version of the TOEFL until 1998, followed by the launch of the TOEFL iBT in 2005. Similarly, although the IELTS consortium

conducted localized, computer-delivered equivalence trials between 2001 and 2005 ([Bruce, 2023](#); [Weir et al., 2007](#)), a remote, unproctored testing environment remained unavailable until the 2020 deployment of IELTS Indicator. This latter platform was introduced strictly as an emergency, interim measure necessitated by global administrative disruptions during the COVID-19 pandemic ([British Council, 2020](#)).

By contrast, the DET bypassed these iterative, legacy structural phases, operating as a natively digital, asynchronous, at-home assessment from its inception in 2016. While its rapid market penetration reflected broader shifts in global technological infrastructure and candidate accessibility rather than an immediate shift in institutional construct preference, the macro-environmental disruptions of 2020 permanently altered the instrument's trajectory. The COVID-19 pandemic served as a profound market catalyst; as physical testing centers shuttered worldwide, the DET's established, cloud-based remote infrastructure remained entirely operational.

This systemic accessibility precipitated an unprecedented 2000% surge in candidate volume during the 2020 calendar year, concurrent with the rapid onboarding of approximately 2,000 new adopting institutions ([Issuu, 2022](#)). As traditional testing vectors plummeted or froze, global candidate participation shifted toward remote alternatives, causing annual DET volumes to skyrocket from an estimated 15,000 test-takers in 2019 to 300,000 in 2020, and ultimately climbing to approximately 700,000 candidates by 2024 ([Arno, 2026](#)). This sustained longitudinal momentum indicates that the DET has evolved past its initial categorization as a pandemic-era administrative stopgap, instead solidifying its position as a permanent and formidable challenger to the historically entrained TOEFL and IELTS duopoly.

Nevertheless, macro-level operational scale remains a critical point of divergence within the international testing landscape. Although, the early historical adoption rates of the IELTS pale in comparison to the rapid digitally leveraged expansion of the DET, traditional examinations continue to command dominant market share. For context, the IELTS maintained its global hegemony with annual registrations exceeding 4.1 million candidates in 2024 ([Market Reports World, 2024](#)). This stark volumetric difference underscores that while digital-first instruments have democratized access and restructured candidate choices, legacy assessments still retain a massive, highly institutionalized global footprint.

The above growth patterns can also be seen at the local level through internal admissions data from the University of Minnesota obtained by the researcher. Table 5 details the volume of various proficiency test submitted over the last five academic years:

Table 5:
Count of U of M English Proficiency Submissions by Test Type and Academic Year, 2020/21-2024/25

English Proficiency Tests	2020-21	%	2021-22	%	2022-23	%
TOEFL	7,145	70.8%	7,222	59.1%	7,008	51.0%
IELTS	2,491	24.7%	3,562	29.2%	4,707	34.3%
DET	407	4.0%	1,375	11.3%	1,919	14.0%
ITOFL	48	0.5%	54	0.4%	94	0.7%
Grand Total	10,091		12,213		13,728	
English Proficiency Tests	2023-24	%	2024-25	%	Grand Total	%
TOEFL	7,463	48.2%	6,423	47.9%	35,261	54.3%
IELTS	5,412	35.0%	4,668	34.8%	20,840	32.1%
DET	2,476	16.0%	2,217	16.5%	8,394	12.9%
ITOFL	122	0.8%	105	0.8%	423	0.7%
Grand Total	15,473		13,413		64,918	

Note: ITOFL stands for Institutional Test of English as a Foreign Language

The above data was provided by University of Minnesota's Office of University Data and Institutional Reporting ([University of Minnesota, 2025](#)). Although the university noted that the 2024-25 figures currently reflect a reporting lag, the majority of submissions remain TOEFL-based. This predominance is expected, given that TOEFL was developed in the United States and has historically served as the primary assessment for domestic admissions. Notably, while TOEFL submissions remained relatively stable from 2020 to 2024, both IELTS and DET recorded year-over-year increases.

Despite an overall decline in total submissions for the 2024-25 cycle (attributed largely to the aforementioned data lag), the DET's market share reached a milestone, accounting for over 15% of total submissions for the first time. This represents a significant increase from just 4% only four years prior. Such rapid institutional adoption suggests that the DET is a formidable competitor that warrants rigorous academic scrutiny alongside established benchmarks like TOEFL and IELTS.

However, increasing volume and popularity do not, by themselves, confirm a test's validity. To evaluate and compare the underlying quality of TOEFL, IELTS and DET, one must first establish the fundamental purpose and requirements of a high-stakes proficiency assessment.

Purpose of Proficiency Tests

According to the American Council on the Teaching of Foreign Languages (ACTFL), language proficiency "specifically describes one's ability to use a language for real-world purposes to accomplish real-world linguistic tasks across a wide range of topics and settings." (Understanding Language Proficiency section) In this context, a proficiency assessment allows candidates to demonstrate the practical application of their linguistic knowledge. Highly

proficient users must remain accurate and adaptable across both conversational and professional domains ([Language Testing International, n.d.](#)). Consequently, a proficiency test evaluates practical communication skills in situational contexts rather than rote memorization or vocabulary breadth ([Tomedes, n.d.](#)).

When candidates sit for assessments like TOEFL, IELTS or DET, their objective is to meet specific score thresholds established by their target institutions. These scores serve as a proxy for linguistic readiness, indicating whether a student possesses the English competency deemed necessary for academic admission. For example, at the University of Minnesota, admission typically requires a minimum score of 79 on the TOEFL, 6.5 on the IELTS or 110 on the DET. The underlying assumption is that students reaching these benchmarks will achieve linguistic success within a collegiate environment. However, correlating a single test score with future academic performance remains a complex challenge.

In establishing these admissions ‘cutoffs,’ researchers often seek to identify connections between test scores and academic outcomes. Yet, as [Graham \(1987\)](#) notes, these correlations are notoriously difficult to define, primarily because ‘language proficiency’ and ‘academic success’ lack universally accepted definitions. If proficiency is defined by a score, the question remains: which assessment is the most accurate metric, particularly when exams as distinct as the TOEFL, IELTS and DET employ different scoring rubrics? Furthermore, the definition of academic success is equally fluid. While studies utilize grade point average (GPA) as a primary metric, this can be problematic; GPA does not account for course load density or the varying linguistic demands of different academic disciplines.

Despite these conceptual hurdles, research has continued across several decades. [Graham’s \(1987\)](#) review of literature from the 1960s through 1980s reveals a spectrum of

inconsistent findings. Some researchers reported negative results—suggesting that low proficiency did not necessarily preclude success—while others found mixed results, where variables such as field of study or student level (graduates vs. undergraduate) mitigated the impact of language scores. Even in studies showing positive correlations the diversity of participants and methodologies made broad generalizations nearly impossible. Ultimately, [Graham \(1987\)](#) concluded that the relationship between proficiency and success remains “murky indeed.” (p. 506).

Contemporary research mirrors these historical inconsistencies. [Ihlenfeldt & Rios \(2022\)](#) conducted a comprehensive meta-analysis synthesizing 32 primary studies published after 2005—the pivotal period following the widespread transition of the TOEFL and IELTS to internet-based testing platforms. In reviewing prior literature, they noted that primary research has historically demonstrated “markedly mixed results when predicting academic success from English admissions exams” (p. 278). By aggregating 132 unique effect sizes while accounting for confounding variables such as institutional control type (public versus private) and publication bias, they ultimately identified a consistent, albeit weak, overall positive correlation ($r = .231$) between standardized English test scores and collegiate GPA. Furthermore, they emphasized that the predictive validity of these language assessments is considerably weaker than that of other traditional academic metrics, such as the American College Testing (ACT), Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT), or high school GPA, signaling that language proficiency scores capture only a narrow sliver of a student’s broader academic capability.

Collectively, these findings suggest a common theme: while large-scale proficiency tests possess a degree of predictive validity, their utility is limited. The variation across decades of research highlights the fact that no single test can serve as a definitive predictor of student

outcomes. Both [Graham \(1987\)](#) and [Ihlenfeldt and Rios \(2022\)](#) conclude with a similar caution: admissions departments should avoid over-reliance on proficiency scores, as success at the university level is influenced by a multitude of factors beyond linguistic competence.

Validity in Proficiency Testing

The issues previously discussed directly address the concept of *test validity*. This property is fundamental to ensuring that an assessment produces meaningful and actionable results for its stakeholders. Broadly defined, validity is "the degree to which a test measures what it requires, or intends, to be measuring" ([Brown, 1996](#), p. 231). From the perspective of Cambridge University, a test achieves validity when it serves as an authentic representation of real-life language use ([Cambridge English, 2025](#)). Consequently, an English proficiency test designed for university admission must accurately reflect the specific linguistic demands a student will encounter in an academic environment.

In the field of test development, validity is a multifaceted construct. [Hashemi and Daneshfar \(2018\)](#) identify three foundational types: *construct*, *content*, and *face validity*. Nested within construct validity is *criterion-related validity*, which assesses how a new instrument compares against an established external benchmark. A critical subset of this is *predictive validity*—the degree to which a test score can forecast a candidate's future performance. It is this specific metric that researchers utilize to determine if high scores on the TOEFL, IELTS or DET truly correlate with subsequent academic success ([Hashemi & Daneshfar, 2018](#)).

A robust correlation between test scores and student outcomes is a hallmark of a reliable proficiency assessment. Whether for the test-taker—who seeks assurance that they possess the linguistic aptitude to thrive in higher education—or the institution—which must verify the caliber of its international applicants—the validity of the assessment remains of paramount importance.

Predictive Validity of TOEFL

[Wongtrirat \(2010\)](#) conducted a meta-analysis synthesizing 22 empirical studies published between 1987 and 2009 to examine the relationship between TOEFL scores and the academic achievement of international students matriculated at higher education institutions in the United States. Academic achievement was operationalized through cumulative GPA or course completion rates, with primary analysis utilized correlation coefficients (r) transformed into standardized effect sizes to evaluate predictive validity.

Across the 22 aggregated studies the meta-analytic sample comprised a total of 3,937 undergraduate and graduate international students. Although the underlying dataset exhibited high statistical heterogeneity with a wide distribution of primary correlation values, the aggregated mean effect size indicated that TOEFL performance yields a comparatively modest predictive capacity regarding student GPAs (Fisher's $Z = .181$). Furthermore, moderator analysis revealed no statistically significant differences in this predictive ability when comparing undergraduate and graduate cohorts ([Wongtrirat, 2010](#)).

A parallel meta-analytic synthesis was executed to evaluate the correlation between admission-level TOEFL scores and subsequent academic course completion metrics. Drawing upon a sub-sample of 1,358 international students, this secondary analysis yielded a nearly identical baseline effect size, establishing that TOEFL scores possess a small predictive relationship with course completion rates (Fisher's Z value - $.173$), with no observed variance across academic levels.

Crucially, [Wongtrirat \(2010\)](#) posited that course completion may theoretically offer a superior reflection of systematic academic achievement relative to GPA, given that international

student visa mandates legally enforce full-time institutional enrollment. However, a structural limitation of the extant literature was noted; while all 22 selected studies reported GPA metrics, only six explicitly integrated course completion data. Ultimately, the restricted linear correlation between TOEFL benchmarks and academic success is attributed to extraneous construct variance—specifically the vast array of non-linguistic academic, psychological, and socio-cultural variables that influence international student performance. Consequently, [Wongtrirat \(2010\)](#) cautioned institutions against isolating standardized language metrics during selection processes, instead advocating for holistic admissions frameworks that complement test scores with secondary predictive markers, such as academic records, recommendation letters, and structured interviews.

[Dabaliz et al. \(2018\)](#) investigated the predictive validity of pre-admission variables among 737 medical students across all academic years (years 1-6) at an English-medium Saudi Arabian university. The pre-admission metrics evaluated included cumulative high school averages, National Achievement Test (NAT) scores, General Aptitude Test (GAT) scores, and performance on the TOEFL and IELTS. These variables were assessed to determine their predictive relationship with cumulative Grade Point Average (cGPA), and progress test (PT) scores across preclinical (years 1-3) and clinical (years 4-6) phases, with statistical significance established at $p < 0.05$.

Statistical analysis revealed that both NAT and TOEFL were significant positive predictors of student' cGPA during the preclinical years, whereas GAT was a significant negative predictor; cumulative high school averages and IELTS scores were not statistically significant. Conversely, during the clinical years, none of the pre-admission variables served as significant predictors of cGPA. Furthermore, none of the evaluated variables significantly predicted PT performance across either academic phase.

[Dabaliz et al. \(2018\)](#) hypothesized that the TOEFL assessment predicted preclinical cGPA because non-native speakers face initial linguistic barrier upon entering an English-medium university; however, by the fourth year, students have typically acclimated to the language of instruction. Additionally, the preclinical curriculum heavily emphasizes reading and writing, whereas the clinical years prioritize verbal communication and clinical proficiency. The researchers supported this rationale by citing [Roemer \(2002\)](#), who concluded that the TOEFL is primarily accurately predicts grammar and reading proficiency rather than broader communicative competence..

Collectively, these findings indicate that while the TOEFL possesses some predictive utility regarding initial academic success, this ability is inherently limited and should not serve as the sole criterion for admission. Consistent with this limitation, ETS, the developer of the TOEFL, explicitly recommends that institutions avoid using rigid cut-off scores for student selection ([Dabaliz et al., 2018](#); [Tannenbaum & Wylie 2008](#)).

Predictive Validity of IELTS

[Dang and Dang \(2021\)](#) investigated the predictive validity of the IELTS on academic performance among a cohort of eighty Vietnamese students across thirty-one UK universities. The study utilized an online questionnaire followed by semi-structured interviews with two undergraduate and two postgraduate students to explore how IELTS scores and preparatory courses influenced subsequent academic success.

Correlation analyses between the participants' IELTS scores and first-semester GPA revealed that while both overall scores and sub-skills were statistically significant predictors, the correlations remained weak. Interview data supported these findings; while students

acknowledged the importance of IELTS skills for their studies, they noted a disconnect between the test and collegiate reality. Specifically, although students identified writing as the most critical skill, it demonstrated the weakest correlation in the quantitative data. Participants noted that university-level essay writing was significantly more complex than IELTS writing subtests, requiring greater information synthesis and more sophisticated vocabulary.

Ultimately, while [Dang and Dang's \(2021\)](#) found a positive correlation—suggesting that higher scores generally track with better performance—the weakness of that correlation implies that passing the exam does not guarantee academic success. Consequently, they argue that while IELTS scores are useful for assessing language proficiency, they should not serve as the main criterion for admissions decisions.

To address the limitations of a single-ethnic study, [Barkaoui \(2025\)](#) expanded the research scope to include 6,481 undergraduates at a Canadian university. This large-scale study examined the associations between entry-level English proficiency scores, admission GPA, first-semester GPA, and longitudinal GPA trajectories over ten semesters. The analysis also controlled for test type and the student's major, categorized into Social Sciences & Humanities (SSH), Business & Economics (BE), and Hard & Life Sciences & Engineering (HLSE).

The findings revealed that for IELTS test-takers, the overall score had a significant positive association with first-semester performance, with an average GPA increase of 0.5 for every one-band increase on the IELTS. Furthermore, higher IELTS scores were associated with a more stable GPA over time, exhibiting less decline across ten semesters compared to lower-scoring peers. However, the strength of this association varied by discipline, with the strongest correlation found in SSH and the weakest in BE.

Intriguingly, [Barkaoui \(2025\)](#) noted that these predictive associations were test-dependent. While IELTS scores showed significant positive associations with first-semester GPA across all majors, TOEFL scores demonstrated a weak and sometimes negative association. The researcher observed that, “It is not clear why the IELTS has a stronger association with Semester 1 GPA compared to the TOEFL iBT” (p. 271), highlighting a need for further investigation into the structural differences between the two assessments.

As evidenced by the research summarized here, while there is evidence supporting the predictive validity of the IELTS, it is far from comprehensive. These studies highlight how different demographics and variables—such as major or specific linguistic tasks—can alter outcomes. Ultimately, despite decades of scrutiny “research on the predictive validity of IELTS on subsequent academic performance has been inconclusive” ([Dang & Dang, 2021](#), p. 84).

Predictive Validity of DET

In contrast to IELTS, the DET’s research base is less developed, due to its shorter operational history. However, its rapid ascension during the COVID-19 pandemic and continued growth have prompted a new wave of predictive validity studies.

[Isaacs et al., \(2023\)](#) conducted a large-scale study involving 1,881 DET test-takers at a UK university during the pandemic. To provide context, they compared these results against 3,087 students who entered during the same period using TOEFL or IELTS scores. Their findings indicated that DET scores positively correlated with first-year grades across the entire sample, with postgraduate (PGT) students showing the strongest correlation. Interestingly, while students in Physical Sciences and Engineering (PSE) entered with lower DET scores, they ultimately achieved higher first-year grades than their peers in other disciplines. Regarding

nationality, non-Chinese postgraduate students showed the strongest positive association, while Chinese students—who utilized the DET in higher proportions—recorded significantly lower test scores and academic grades.

When comparing the three assessments, Isaacs et al. found that at the postgraduate level, all three tests positively correlated with grades, though the correlations for IELTS and TOEFL were stronger. At the undergraduate level, results were more fragmented; TOEFL maintained a weak positive correlation, while both IELTS and DET exhibited weak negative correlations. The researcher attributed these undergraduate discrepancies to high sample heterogeneity rather than inherent test flaws.

To normalize the data across different scoring scales, the researchers mapped scores to the Common European Framework of Reference for Languages (CEFR)—an international standard ranging from A1 (Basic) to C2 (Proficient) ([Cambridge University Press & Assessment, n.d.](#)).

At the B2 and C1 levels, DET takers generally achieved lower university grades than their IELTS counterparts. However, at the C2 (highest) level, postgraduate DET takers outperformed those who took traditional exams. [Isaacs et al., 2023](#) suggested that these discrepancies might stem from differences in scoring methodologies or the DET's role as an accessible alternative that may cater to students from varied socioeconomic backgrounds.

Conversely, [Tabassum \(2023\)](#) reported markedly different results in a study of 1,460 student records at a Canadian university. In this instance, the DET demonstrated the strongest correlation with first-semester GPA—outperforming both TOEFL and IELTS. While DET's predictive power weakened by the second semester, it remained statistically significant,

whereas the TOEFL's correlation lost significance entirely. Tabassum's findings suggest that a higher DET score correlates more clearly with initial GPA, though the long-term predictive accuracy remains limited for all assessments.

Expanding into the post-pandemic era, [Isbell and McGehee \(2025\)](#) studied 397 international students at the University of Hawai'i. Their results highlighted the complexity of these relationships; the majority of correlations between test scores and first-year GPA were not statistically significant, with the exception of graduate-level IELTS scores and scores from conditionally admitted students. Despite the small sample size, the researchers concluded that the DET functions as a legitimate proficiency assessment alongside the IELTS and TOEFL, even if it displays weaker correlations in specific subgroups.

Synthesis of Findings and Research Direction

Whether TOEFL, IELTS or DET, English proficiency test research presents a mosaic of mixed results. While the DET occasionally exhibits weaker correlations than its predecessors, a consistent positive relationship between scores and success persists across various studies. The limitations of these studies—ranging from demographic specificity to small sample size—underscore the need for large-scale, longitudinal research as DET adoption continues to climb.

To date, the majority of quantitative research has relied on GPA as a proxy for academic success. While scholars generally accept GPA as a sufficient metric ([Graham, 1987](#); [Ihlenfeldt & Rios, 2022](#)), the extreme variability in these predictive validity studies suggests that a numerical correlation alone is insufficient. To provide a more holistic perspective, the present research moves beyond the traditional test score/GPA nexus. By gathering qualitative data directly from test-takers regarding their experiences and perceptions of linguistic readiness, this

study seeks to add a necessary human dimension to the debate over English proficiency and academic success.

Methodology

The current study utilizes a test-taker's English proficiency scores—both composite and subscores—as the primary baseline for analysis. Participants are asked to self-report their level of confidence across the four fundamental linguistic domains: reading, writing, speaking, and listening. In this study, they were not asked to provide GPA as this metric emphasizes a quantitative connection with proficiency scores, rather than test-takers' perception of their confidence and preparation. This study hypothesizes a positive relationship¹ between objective proficiency scores and subjective confidence levels; theoretically, these combined factors should serve as a stronger indicator of a student's academic potential. Conversely, a significant divergence between these metrics—such as a high test score paired with low self-reported confidence—may suggest that the assessment lacks the predictive accuracy necessary to gauge a student's readiness for collegiate-level rigors.

Furthermore, the study examines the perceived accuracy of these scores from the test-taker's perspective. If a student perceives the test as a highly accurate reflection of their linguistic ability, their score is expected to demonstrate a stronger relationship with their actual academic performance. In contrast, if students report low perceived accuracy, the score's utility as a predictor of academic success may be diminished, regardless of the numerical result.

¹ To maintain methodological clarity, this study employs the term "*relationship*" rather than "*correlation*". This distinction ensures the findings are not conflated with strict statistical correlation, aligning with the study's mixed-methods approach.

Study Research Questions

Grounded in the aforementioned framework of test validity and student self-efficacy, this study addresses the following research questions:

- **RQ1: What relationships exist between a test-taker's objective proficiency scores and their perceived linguistic confidence regarding their readiness for collegiate academic study?**
- **RQ2: To what extent do proficiency test scores align with a test-taker's perception of the assessment's accuracy in measuring their academic language preparation?**
- **RQ3: How do perceptions of test accuracy and linguistic readiness differ among test-takers of the IELTS, DET, and TOEFL?**

Given the distinct developmental histories of the TOEFL, IELTS and DET—and the currently inconsistent data regarding their predictive validity—this study investigates the intersection of student perception and standardized testing results. By exploring how candidates experience these assessments, this research aims to determine if score-based benchmarks adequately reflect a test-taker's own sense of academic readiness. Ultimately, these insights aim to deepen institutional understanding of English proficiency metrics, highlighting that standardized scores offer only a partial glimpse of student readiness and ensuring that university support systems that can more effectively prepare students for the rigors of higher education.

Research Design

A student survey was developed (see Appendix A) to assess test-takers' attitudes regarding English proficiency assessments. The instrument was structured into three distinct sections. *Part 1* evaluated the participants' linguistic background prior to their arrival in the United States, including years of English language instruction, their intended level of study (undergraduate or graduate), and whether they were required to enroll in supplemental English

courses upon matriculation. The objective of this section was to establish a baseline of each participant's language preparation and its potential influence on their academic readiness.

Part 2 required participants to report their English proficiency test scores, including composite results and individual subscores. In addition to quantitative data, this section gauged participants' perceptions regarding the accuracy of their scores. Test-takers were also provided with an open-ended format to offer qualitative commentary on the alignment between their scores and their perceived readiness for academic study.

Finally, the survey invited participants to volunteer for a follow-up semi-structured interview. These interviews were designed to elicit deeper insights into the testing process and how the experience influenced the individuals' outlook on their subsequent academic trajectory (see Appendix B for the interview protocol).

Study Procedure

This study was conducted during the fall semester of 2025. Because the research involved human participants via online survey, a formal inquiry was submitted to the Institutional Review Board (IRB), which categorized the study as exempt (IRB ID: STUDY00026663). Following this determination, the survey was distributed through several recruitment channels to maximize reach. First, the survey link and description were posted on the researcher's social media accounts. Second, the survey was sent directly to university peers and classmates via email or text. Finally, recruitment information was disseminated through campus organizations and departments at the University of Minnesota that specifically serve international student and language populations.

Because this study examines the perceptions and experiences of English proficiency testing in relation to higher education success, the eligibility criteria required participants to have taken a proficiency exam for university admission and to provide their overall scores and subscores for analysis. However, respondents were not required to specify whether they were first-year, first-semester students, which introduces a limitation regarding the exact academic timeline of the sample.

The survey was housed in a Google Form, which respondents could reach via an online link. It was open for seven weeks, during which a total of thirty-five responses were received. At the end of the survey (Part 3), respondents were asked if they were willing to participate in a follow-up interview with the researcher. Only those willing were asked for their email address to facilitate further contact for a potential interview.

From the pool of survey volunteers, two participants were selected using purposive sampling to ensure the data closely aligned with study's specific research objectives ([Laerd Dissertation, 2012](#)). Because the test-takers' DET experience is a central, under-researched component of this study, the selection criteria prioritized respondents who had completed that specific assessment. Upon reviewing the survey responses, two DET test-takers were identified who met these criteria and agreed to participate. Conducted via Zoom in early December 2025, the interviews utilized a semi-structured protocol that prompted participants to share insights regarding their English language acquisition, proficiency testing experiences, and initial transitions to collegiate academics (see Appendix B).

Data Preparation

Following the classification model established by [Isaacs et al., \(2023\)](#) this study categorized academic majors into three domains: Life Sciences (LS), Physical Sciences and

Engineering (PSE), and Social Science and Humanities (SSH). Upon reviewing the participant data, only one respondent (“Biostatistics”) fell within the LS category. To ensure statistical viability, the LS and PSE groups were consolidated into a single category: Life, Physical Science, and Engineering (LPSE). This consolidation was predicated on the assumption that the linguistic demands of biostatistics align more closely with the technical requirements of the hard sciences than with those of the humanities. Consequently, the final analytical sample consisted of 23 majors in SSH (71.9%) and 10 in LPSE (28.1%); one “undecided” entry was excluded from the analysis. See Table 6 below for a listing of all majors represented and their categorization:

Table 6:

Test-taker Majors and Categorizations

Category	Major	Category	Major
SSH	PhD Arts in Education	LPSE	Biostatistics
SSH	Curriculum and Instruction	LPSE	mechanical engineering
SSH	cineme	SSH	Multilingual Education
SSH	Multilingual education	SSH	TESOL
SSH	Science Education	LPSE	cege
LPSE	Mechanical engineering	SSH	Curriculum & Instruction
LPSE	Computer Science	SSH	Multilingual Education
SSH	STEM education	SSH	Multilingual Education
SSH	TESOL	SSH	Masters in Public Policy
SSH	Educational Psychology	SSH	Elementary Education
	Undecided	SSH	Mass communication

LPSE	Physics	SSH	Applied Economics
SSH	comparative education	SSH	Learning Technologies
LPSE	Chemical Engineering	SSH	TESL
LPSE	Electrical and computer engineering	SSH	mass communication
LPSE	Math+CS	SSH	Public Policy
SSH	Multi Ed		
SSH (n = 23)		LPSE (n = 10)	

Once the data were cleaned and categorized, the researcher performed a series of descriptive and comparative analyses based on the following variables:

- Linguistic Level: CEFR equivalents (B1, B2, C1, C2)
- Self-Efficacy: Confidence levels (Binary: 1 for confident, 0 for not confident)
- Assessment Type: Proficiency test taken (DET, IELTS, TOEFL)
- Academic Discipline: Major Category (SSH, LPSE)
- Supplemental Instruction: Requirement for post-admission U.S. English courses (Binary: Yes/No)
- Perceived Accuracy: Likert-scale ratings (1-5 rating for composite scores; qualitative ratings for subscores)
- Preparatory Utility: Perceived ability of the test to prepare the student for study (1-5 rating)

These variables were cross-referenced in various combinations to identify emergent trends within the dataset. While 17 distinct variable combinations were analyzed, the findings highlighted below focus primarily on the relationship between proficiency scores and test-taker perceptions of confidence and accuracy (see Appendix C for a comprehensive data summary).

Results

Confidence and Test Scores

As previously noted, 33 test-taker responses were utilized for the score distribution analysis. After normalizing the results from the three proficiency assessments to their CEFR equivalents, the data revealed that the vast majority of composite scores (97%) were situated within B2-C1 range (CEFR scores span from A1 to C2, with C2 representing the highest level of proficiency).

This concentration was also evident across all linguistic sub-domains, where at least 75% of subscores fell within the B2-C1 bracket. Specifically, the receptive skills (Reading and Listening) exhibited higher concentrations at the C1 level, whereas the productive skills (Writing and Speaking) were more densely clustered at the B2 level. These distributions are detailed in Table 7 below. *Note: Bolded percentages indicate the majority distribution within each respective category.*

Table 7:

CEFR Scores Overall and Each Skill

Score	Overall		Reading		Writing		Speaking		Listening	
C2	1	3.0%	6	18.2%	5	15.2%	4	12.1%	8	24.2%
C1	20	60.6%	20	60.6%	11	33.3%	6	18.2%	16	48.5%
B2	12	36.4%	7	21.2%	16	48.5%	22	66.7%	9	27.3%
B1	0	0.0%	0	0.0%	1	3.0%	1	3.0%	0	0.0%

Participants were asked to identify the specific linguistic domains in which they felt most proficient at the onset of their university studies. The survey results indicate that students reported the highest levels of self-efficacy in *Reading* (81.8%), while *Speaking* represented the

area of lowest confidence (36.4%). These findings suggest a notable disparity between receptive and productive skill confidence among incoming international students. The comprehensive results are detailed below:

Table 8:

Confidence Level for Each Skill

	Reading		Writing		Speaking		Listening	
Confident	27	81.8%	18	54.5%	12	36.4%	23	69.7%
Not Confident	6	18.2%	15	45.5%	21	63.6%	10	30.3%

When the CEFR subscores are cross-referenced with the reported confidence levels for each specific skill (excluding aggregate composite scores), the data demonstrate a high degree of alignment. For instance, the combined percentage of participants scoring at C1 and C2 levels in *Reading* (78.8%) closely mirrors the reported confidence level for that domain (81.8%), a marginal difference of only 3%. This consistency persists across all objective score concentration and subjective confidence.

While the previous table illustrated aggregate overall confidence levels, it was also necessary to examine how these perceptions fluctuated across the CEFR spectrum from B1 to C2. By disaggregating the data into the four primary language skills, the researcher sought to determine the extent to which incremental score increases directly influenced test-taker self-efficacy. The results of this distribution analysis are presented below:

Table 9:

CEFR Scores and Confidence Level

	Reading		Writing		Speaking		Listening	
C2 / Conf	6	100.0%	3	60.0%	3	75.0%	6	75.0%
C2 / N Conf	0	0.0%	2	40.0%	1	25.0%	2	25.0%

C1 / Conf	18	90.0%	8	72.7%	3	50.0%	10	62.5%
C1 / N Conf	2	10.0%	3	27.3%	3	50.0%	6	37.5%
B2 / Conf	3	42.9%	6	37.5%	6	27.3%	7	77.8%
B2 / N Conf	4	57.1%	10	62.5%	16	72.7%	2	22.2%
B1 / Conf	0	0.0%	1	100.0%	0	0.0%	0	0.0%
B1 / N Conf	0	0.0%	0	0.0%	1	100.0%	0	0.0%

The data indicate that, in most instances, higher CEFR levels are related to an increased likelihood of linguistic confidence. The notable exception is the *B2 Listening* subscore, where confidence levels deviated from the established trend. It should be noted that because only two participants scored at the *B1* level across all skill domains, these specific percentages may lack statistical reliability and should be interpreted with caution.

Three distinct English proficiency assessments were represented in the survey, with the following distribution: TOEFL (n=20, 60.6%), IELTS (n=10, 30.3%), and DET (n=3, 9.1%). Despite the relatively small sample size, this distribution is largely representative of the actual admissions data reported by the University of Minnesota over the past five years (TOEFL: 54.4%; IELTS: 31.3%; DET: 13.5%). However, a limitation of the survey is that it did not record the specific year of matriculation for each respondent.

To further investigate these trends, the overall CEFR results and their corresponding subscores were disaggregated by assessment type. The comparative distribution is detailed in the table below:

Table 10:
CEFR Scores By Test

DET (n=3)						
Score	Overall	Reading	Writing	Speaking	Listening	

C2	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
C1	1	33.3%	2	66.7%	0	0.0%	1	33.3%	1	33.3%
B2	2	66.7%	1	33.3%	2	66.7%	2	66.7%	2	66.7%
B1	0	0.0%	0	0.0%	1	33.3%	0	0.0%	0	0.0%
IELTS (n=10)										
Score	Overall		Reading		Writing		Speaking		Listening	
C2	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	20.0%
C1	6	60.0%	7	70.0%	4	40.0%	2	20.0%	5	50.0%
B2	4	40.0%	3	30.0%	6	60.0%	8	80.0%	3	30.0%
B1	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
TOEFL (n=20)										
Score	Overall		Reading		Writing		Speaking		Listening	
C2	1	5.0%	6	30.0%	5	25.0%	4	20.0%	6	30.0%
C1	13	65.0%	11	55.0%	7	35.0%	3	15.0%	10	50.0%
B2	6	30.0%	3	15.0%	8	40.0%	12	60.0%	4	20.0%
B1	0	0.0%	0	0.0%	0	0.0%	1	5.0%	0	0.0%

The data suggest that DET scores are primarily concentrated within the B2 range, with the exception of the reading subscore; however, the limited sample size necessitates a cautious interpretation of these trends. A comparative analysis of IELTS and TOEFL reveals similar distributional patterns across most domains, with the majority of scores clustering within the same CEFR score ranges for both assessments. Notably, the TOEFL sample exhibits a significantly higher frequency of C2-level scores compared to the IELTS.

This comparative assessment serves as an evaluation of *concurrent validity*– the extent to which a new test or alternative assessment aligns with an established criterion ([Hashemi & Daneshfar, 2018](#)). In this instance, while IELTS and TOEFL demonstrate a consistent range of score representation, the DET data show a distinct divergence. While the current sample size is insufficient to conclude that the DET lacks concurrent validity, these results emphasize the

necessity for larger-scale research to more accurately define DET's psychometric characteristics in relation to traditional benchmarks.

The subsequent phase of this analysis examines linguistic confidence levels across the four domains, disaggregated by assessment type. The findings are detailed below:

Table 11:

Confidence Level for Each Skill by Test

DET (n=3)								
	Reading		Writing		Speaking		Listening	
Confident	3	100.0%	2	66.7%	1	33.3%	2	66.7%
Not Confident	0	0.0%	1	33.3%	2	66.7%	1	33.3%
IELTS (n=10)								
	Reading		Writing		Speaking		Listening	
Confident	7	70.0%	5	50.0%	3	30.0%	7	70.0%
Not Confident	3	30.0%	5	50.0%	7	70.0%	3	30.0%
TOEFL (n=20)								
	Reading		Writing		Speaking		Listening	
Confident	17	85.0%	11	55.0%	8	40.0%	14	70.0%
Not Confident	3	15.0%	9	45.0%	12	60.0%	6	30.0%

The findings indicate that all three assessments exhibit a consistent trend in participant self-efficacy: test-takers reported the highest levels of confidence in their *Reading* ability and lowest in their *Speaking* ability. These results prompt a critical inquiry: do higher CEFR subscores on each respective test relate to increased linguistic confidence? To investigate this, confidence levels were disaggregated by CEFR subscore and categorized by proficiency assessment. The results of this relational analysis are presented below:

Table 12:*CEFR Scores and Confidence Level by Test*

DET (n=3)								
	Reading		Writing		Speaking		Listening	
C2 / Conf	0	0.0%	0	0.0%	0	0.0%	0	0.0%
C2 / N Conf	0	0.0%	0	0.0%	0	0.0%	0	0.0%
C1 / Conf	2	100.0%	0	0.0%	0	0.0%	0	0.0%
C1 / N Conf	0	0.0%	0	0.0%	1	100.0%	1	100.0%
B2 / Conf	1	100.0%	1	50.0%	1	50.0%	2	100.0%
B2 / N Conf	0	0.0%	1	50.0%	1	50.0%	0	0.0%
B1 / Conf	0	0.0%	1	100.0%	0	0.0%	0	0.0%
B1 / N Conf	0	0.0%	0	0.0%	0	0.0%	0	0.0%
IELTS (n=10)								
	Reading		Writing		Speaking		Listening	
C2 / Conf	0	0.0%	0	0.0%	0	0.0%	1	50.0%
C2 / N Conf	0	0.0%	0	0.0%	0	0.0%	1	50.0%
C1 / Conf	7	100.0%	3	75.0%	1	50.0%	3	60.0%
C1 / N Conf	0	0.0%	1	25.0%	1	50.0%	2	40.0%
B2 / Conf	0	0.0%	2	33.3%	2	25.0%	3	100.0%
B2 / N Conf	3	100.0%	4	66.7%	6	75.0%	0	0.0%
B1 / Conf	0	0.0%	0	0.0%	0	0.0%	0	0.0%
B1 / N Conf	0	0.0%	0	0.0%	0	0.0%	0	0.0%
TOEFL (n=20)								
	Reading		Writing		Speaking		Listening	
C2 / Conf	6	100.0%	3	60.0%	3	75.0%	5	83.3%
C2 / N Conf	0	0.0%	2	40.0%	1	25.0%	1	16.7%
C1 / Conf	9	81.8%	5	71.4%	2	66.7%	7	70.0%
C1 / N Conf	2	18.2%	2	28.6%	1	33.3%	3	30.0%
B2 / Conf	2	66.7%	3	37.5%	3	25.0%	2	50.0%
B2 / N Conf	1	33.3%	5	62.5%	9	75.0%	2	50.0%
B1 / Conf	0	0.0%	0	0.0%	0	0.0%	0	0.0%
B1 / N Conf	0	0.0%	0	0.0%	1	100.0%	0	0.0%

Due to the high level of data granularity and the limited sample size for the DET, it was not possible to identify statistically significant trends for that specific assessment. However, the TOEFL data demonstrated a clear trajectory: higher CEFR levels consistently coincided with increased reported confidence. Similar patterns were observed in the IELTS data, with the notable exception of the listening sub-skill. It should also be noted that the IELTS sample yielded only two C2-level scores, rendering the confidence relationship for that specific proficiency tier inconclusive.

In aggregate, these results suggest that higher proficiency scores generally relate to increased linguistic self-efficacy. This relationship is critical, as a positive mindset and perceived readiness are essential components of a student's preparation for the rigors of higher education. This aligns with the broader literature suggesting that while scores are a primary metric, subjective readiness remains an influential variable in academic transition.

Accuracy and Test Scores

Participants were also asked to evaluate the extent to which their proficiency test results accurately reflected their linguistic competence. Specifically, respondents rated the perceived accuracy of their composite scores on a 5-point Likert scale (with 5 representing the highest degree of accuracy). These perceptions were tabulated against the corresponding CEFR levels to identify any patterns between objective performance and subjective validation. The distribution of these ratings is detailed in the table below:

Table 13:

Accuracy Feeling by CEFR

		High Accuracy				Low/Moderate Accuracy					
	n	5		4		3		2		1	
C2	1	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
C1	20	4	20.0%	11	55.0%	4	20.0%	0	0.0%	1	5.0%
B2	12	1	8.3%	4	33.3%	6	50.0%	1	8.3%	0	0.0%
B1	0	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%

As previously noted, the aggregate CEFR scores were heavily concentrated within the B2 to C1 ranges (with only a single respondent at the C2 level and none at B1); consequently, the current sample size precludes a definitive conclusion regarding score-based trends. However, the limited data suggest a subtle positive relationship: higher CEFR levels appear to align with slightly higher levels of perceived accuracy. Furthermore, a clear majority of participants—approximately 64%—validated the assessment’s accuracy, assigning a rating of 4 or 5 on the 5-point scale..

Perceived accuracy was further disaggregated by CEFR subscore across the four linguistic domains. For this categorical analysis, respondents evaluated the accuracy of their individual skill scores as ‘very accurate,’ ‘somewhat accurate,’ or ‘not accurate.’ These results are detailed in the table below:

Table 14:

Accuracy Feeling by CEFR Subscore

	n	Very Accurate		Somewhat Accurate		Not Accurate	
Reading							
C2	6	4	66.7%	2	33.3%	0	0.0%
C1	20	11	55.0%	6	30.0%	3	15.0%
B2	7	1	14.3%	5	71.4%	1	14.3%
B1	0	0	0.0%	0	0.0%	0	0.0%

Writing							
C2	5	1	20.0%	3	60.0%	1	20.0%
C1	11	4	36.4%	7	63.6%	0	0.0%
B2	16	7	43.8%	6	37.5%	3	18.8%
B1	1	0	0.0%	0	0.0%	1	100.0%
Speaking							
C2	4	2	50.0%	1	25.0%	1	25.0%
C1	6	2	33.3%	4	66.7%	0	0.0%
B2	22	5	22.7%	14	63.6%	3	13.6%
B1	1	1	100.0%	0	0.0%	0	0.0%
Listening							
C2	8	4	50.0%	4	50.0%	0	0.0%
C1	16	3	18.8%	8	50.0%	5	31.3%
B2	9	3	33.3%	5	55.6%	1	11.1%
B1	0	0	0.0%	0	0.0%	0	0.0%

Across all linguistic domains and CEFR levels, a minimum of 68% of respondents rated their subscores as at least ‘somewhat accurate.’ However, unlike the aggregate CEFR data, the subscores do not exhibit a linear relationship between higher proficiency levels and increased perceived accuracy. This inconsistency may be attributed to the uneven distribution of scores across the four skills. For instance, the *Speaking* domain recorded 22 scores at the B2 level, whereas the *Reading* domain yielded only seven. This disparity in data density within specific ranges complicates the identification of overarching trends for individual sub-skills.

Beyond the general CEFR analysis, the study specifically examined how perceptions accuracy varied across the three different proficiency assessments. A comparative breakdown for the *DET*, *IELTS*, and *TOEFL* is presented below:

Table 15:

Accuracy Feeling by Test

	High Accuracy				Low/Moderate Accuracy					
	5		4		3		2		1	
DET	0	0.0%	2	66.7%	1	33.3%	0	0.0%	0	0.0%
IELTS	3	30.0%	4	40.0%	2	20.0%	1	10.0%	0	0.0%
TOEFL	2	10.0%	10	50.0%	7	35.0%	0	0.0%	1	5.0%

All three assessments demonstrated relatively high levels of perceived accuracy, with at least 60% of test-takers for each exam assigning a rating of 4 or 5. Proportionally, IELTS candidates reported the highest degree of confidence in the accuracy of their results, followed by DET test-takers, and finally those who took the TOEFL.

To further investigate these perceptions across different proficiency levels, perceived accuracy ratings were cross-tabulated by CEFR score and disaggregated by assessment type. The following table provides a detailed breakdown of these findings:

Table 16:

CEFR Scores and Accuracy Feeling by Test

	High Accuracy				Low/Moderate Accuracy					
	5		4		3		2		1	
DET	5		4		3		2		1	
C2	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
C1	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%
B2	0	0.0%	1	50.0%	1	50.0%	0	0.0%	0	0.0%
B1	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
IELTS	5		4		3		2		1	
C2	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
C1	2	33.3%	3	50.0%	1	16.7%	0	0.0%	0	0.0%
B2	1	25.0%	1	25.0%	1	25.0%	1	25.0%	0	0.0%
B1	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
TOEFL	5		4		3		2		1	
C2	0	0.0%	1	100.0%	0	0.0%	0	0.0%	0	0.0%

C1	2	15.4%	7	53.8%	3	23.1%	0	0.0%	1	7.7%
B2	0	0.0%	2	33.3%	4	66.7%	0	0.0%	0	0.0%
B1	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%

A consistent trend is observable across all three proficiency assessments: higher CEFR levels appear to relate positively with higher levels of perceived accuracy. At this stage of the analysis, the data do not indicate that any single assessment significantly outperforms the others in terms of student validation.

To further explore the relationship between perceived accuracy and assessment type, CEFR subscores were cross-referenced with test-takers' qualitative evaluations of accuracy. These findings were then disaggregated by proficiency tests to identify domain-specific variations. The results are detailed in the following table:

Table 17:

CEFR Subscores and Accuracy Feeling by Test

DET Scores	DET	Very Accurate		Somewhat Accurate		Not Accurate	
n=3	R	1	33.3%	1	33.3%	1	33.3%
	W	1	33.3%	1	33.3%	1	33.3%
	S	1	33.3%	2	66.7%	0	0.0%
	L	1	33.3%	2	66.7%	0	0.0%
IELTS Scores	IELTS	Very Accurate		Somewhat Accurate		Not Accurate	
n=10	R	5	50.0%	4	40.0%	1	10.0%
	W	5	50.0%	5	50.0%	0	0.0%
	S	3	30.0%	7	70.0%	0	0.0%
	L	4	40.0%	6	60.0%	0	0.0%
TOEFL Scores	TOEFL	Very Accurate		Somewhat Accurate		Not Accurate	
n=20	R	10	50.0%	8	40.0%	2	10.0%
	W	6	30.0%	10	50.0%	4	20.0%
	S	6	30.0%	10	50.0%	4	20.0%

	L	5	25.0%	9	45.0%	6	30.0%
Total Scores	33						

In this analysis, all three tests are found to have relatively high accuracy, with the combination of ‘very accurate’ and ‘somewhat accurate’ being at least 66% for all tests and skills. The total of each skill and test accuracy feeling can be seen in the following summary table:

Table 18:

Accuracy Feeling Totals by Subscore and Test

Subscores	Very Accurate		Somewhat Accurate		Not Accurate	
R	16	48.5%	13	39.4%	4	12.1%
W	12	36.4%	16	48.5%	5	15.2%
S	10	30.3%	19	57.6%	4	12.1%
L	10	30.3%	17	51.5%	6	18.2%
DET	4	33.3%	6	50.0%	2	16.7%
IELTS	17	42.5%	22	55.0%	1	2.5%
TOEFL	27	33.8%	37	46.3%	16	20.0%

The data indicate that a substantial majority of participants perceived their subscores to be reliable, with cumulative accuracy ratings (‘very accurate’ and ‘somewhat accurate’) exceeding 81% across all domains. On an aggregate level, the *Reading* subscore was identified as the most accurate measure (with 48.5% reporting it as ‘very accurate’), whereas *Listening* was perceived as the least accurate, with 18.2% of respondents labeling their scores as ‘not accurate’.

When disaggregated by assessment type, the total accuracy ratings reveal a compelling trend not previously observed in the aggregate data. Participants who took the *IELTS* reported

a remarkably high level of perceived accuracy ('very accurate' and 'somewhat accurate' = 97.5%), outperforming both DET (83.3%) and TOEFL (80.1%) by a significant margin. Despite the limitations of the current sample size, this marked disparity in student validation is statistically noteworthy and warrants further longitudinal investigation to determine why IELTS results may align more closely with student self-perceptions than other standardized measures.

Qualitative Insights on Score Accuracy

In addition to quantitative ratings, the survey provided participants with an open-ended opportunity to elaborate on the reasoning behind their accuracy assessments. Although this section was optional, over 50% of respondents provided qualitative commentary. The feedback revealed a spectrum of perceptions, ranging from strong validation of the assessments to significant critiques of their practical ability.

Several participants noted that the tests accurately captured the nuances of their linguistic strengths and weaknesses. One *DET* respondent observed, "I think it was accurate because I could understand and read very well, but when it was me speaking complex things, it got more difficult when I started studying." Similarly, an *IELTS* participant stated, "I was feeling I had some capacity to grow while being adequate at the same time. I believe a 7 in IELTS reflects that." A *TOEFL* examinee credited the test's structure, noting, "It was accurate in that the test included a lot of tasks that required integration of the four language skills."

Conversely, other respondents highlighted discrepancies between test performance and actual academic or personal capabilities. A *DET* participant emphasized the psychological burden of the format, stating, "The tests give extra stress of time limit....Even the native speakers will struggle with this." An *IELTS* respondent pointed to institutional barriers, noting, "As a person who is visually disabled I feel like I did not get the best accommodations."

Furthermore, several *TOEFL* participants critiqued the test's construct, arguing that the focus was overly "academic" and failed to incorporate the "everyday" English necessary for social integration.

These qualitative data suggest that while the majority of test-takers perceive these assessments as generally valid, individual variables—such as psychological pressure, accessibility, and the distinction between academic and colloquial English—significantly influence their perceptions of accuracy. These perceptions, in turn, likely impact a student's confidence during their initial transition to higher education.

Perceived Preparatory Utility

Finally, participants were asked to evaluate the extent to which the process of preparing for and completing the proficiency assessment equipped them for the demands of university-level study.

Table 19:

Test Accuracy Feeling vs. Preparatory Ability

Accuracy to Prep	Quantity	Percentage		Accuracy to Prep	Quantity	Percentage
5 to 5	2	40.0%		2 to 5	0	0.0%
5 to 4	1	20.0%		2 to 4	0	0.0%
5 to 3	0	0.0%		2 to 3	0	0.0%
5 to 2	1	20.0%		2 to 2	1	100.0%
5 to 1	1	20.0%		2 to 1	0	0.0%
4 to 5	3	18.8%		1 to 5	0	0.0%
4 to 4	4	25.0%		1 to 4	0	0.0%
4 to 3	6	37.5%		1 to 3	1	100.0%
4 to 2	3	18.8%		1 to 2	0	0.0%
4 to 1	0	0.0%		1 to 1	0	0.0%
3 to 5	0	0.0%				
3 to 4	3	30.0%				
3 to 3	4	40.0%				
3 to 2	0	0.0%				
3 to 1	3	30.0%				

On an aggregate level, ratings for *perceived preparatory utility* largely mirrored those for *perceived accuracy*; specifically, participants who viewed their assessment as a more accurate reflection of their ability also reported higher levels of perceived preparation. However, this relationship did not extend to objective performance metrics. As illustrated in the table below,

there was no discernible linear relationship between higher CEFR scores and the perceived effectiveness of the test-taking process in preparing students for academic study.

Table 20:

CEFR vs. Preparatory Ability

CEFR to Prep	Quantity	Percentage		CEFR to Prep	Quantity	Percentage
C2 to 5	0	0.0%		B2 to 5	0	0.0%
C2 to 4	0	0.0%		B2 to 4	4	33.3%
C2 to 3	0	0.0%		B2 to 3	4	33.3%
C2 to 2	1	100.0%		B2 to 2	2	16.7%
C2 to 1	0	0.0%		B2 to 1	2	16.7%
C1 to 5	5	25.0%		B1 to 5	0	0.0%
C1 to 4	4	20.0%		B1 to 4	0	0.0%
C1 to 3	7	35.0%		B1 to 3	0	0.0%
C1 to 2	2	10.0%		B1 to 2	0	0.0%
C1 to 1	2	10.0%		B1 to 1	0	0.0%

Comparative analysis revealed no significant advantage for any specific assessment regarding its perceived preparatory utility; ratings remained relatively consistent across the *DET*, *IELTS*, and *TOEFL*.

Despite the quantitative distribution of these ratings, the qualitative commentary leaned predominantly toward the negative. Respondents frequently critiqued the disconnect between

standardized testing and the linguistic realities of immersion. One participant dismissed the process entirely, describing it as a “waste of money” that provided no practical benefit. Another highlighted the lack of ecological validity in the assessment content, noting: “Language used in exam is different from language used in authentic life context...so many slang or local words [will not appear] in exam.”

Furthermore, a common theme emerged regarding the distinction between genuine proficiency and ‘test-wiseness.’ As one respondent observed, “You just need to learn how to perform on the test, it does not show your real ability, just a picture of your performance.” Collectively, these perspectives suggest that test-takers perceive the preparatory phase as an exercise in mastering test-taking strategies rather than a meaningful opportunity for English language acquisition or academic readiness.

Qualitative Case Studies: Perspectives on the Duolingo English Test (DET)

Due to the small sample of DET test-takers in the quantitative data, the researcher conducted semi-structured follow-up interviews with the two participants who took DET and agreed to be interviewed. This qualitative addition provides deeper insight into the nuances of the DET experience.

Case Study 1: The Experienced Educator (TT1)

Background and Linguistic Development. Test-taker 1 (TT1) completed the DET in late 2021. Raised in Asia, TT1 was exposed to English instruction from primary school through their undergraduate studies in chemical engineering, where English was the medium of instruction. Prior to their U.S. graduate application, TT1’s linguistic journey included three months of intensive coaching for the IELTS, a brief period of study in the United Kingdom, and a year of English for Academic Purposes (EAP) in Australia, where they earned an Upper

Intermediate certificate. Following this, TT1 spent six years as an educator in Singaporean international schools, using English as the primary vehicle for pedagogy and professional communication.

The DET Experience. Despite a decade of professional immersion, TT1 was required to submit a proficiency score for U.S. graduate admission—a requirement they initially met with skepticism. TT1 chose the DET for its convenience. Unlike their rigorous IELTS preparation years prior, TT1’s preparation was minimal, focusing primarily on acclimating to the digital interface and typing requirements. For TT1, the DET was viewed merely as a bureaucratic hurdle rather than a meaningful measure of ability.

Academic Transition. Although TT1’s results fell within the CEFR B2 range—lower than might be expected given their background—they reported no linguistic barriers upon entering graduate school. They attributed their academic success to years of professional immersion rather than the test itself, suggesting that for high-proficiency users, the DET may function as a ‘formality’ that does not fully capture long-term linguistic competence.

Case Study 2: The Practical Learner (TT2)

Background and Linguistic Development. Test-taker 2 (TT2) followed a different linguistic path, attending a bilingual Spanish/English primary school in South America before transitioning to Spanish-only instruction through college. After earning a degree in Visual Arts, they taught in a bilingual setting for five years. TT2 had previously taken the TOEFL twice, but chose the DET for their graduate application to avoid the higher costs and logistical constraints of retaking the TOEFL.

The DET Experience. TT2 utilized the DET’s official preparatory materials and perceived the test as more ‘practical’ and less ‘scholarly’ than the TOEFL. While they felt they

performed better on DET due to its accessible nature, this perception of ‘ease’ created a false sense of security regarding their academic readiness.

Academic Transition and Reflection. Upon matriculation, TT2 experienced a significant ‘culture shock.’ Despite their successful DET score, they struggled to keep pace with native speakers in seminars and found academic reading and writing tasks to be prohibitively time-consuming. TT2 eventually achieved academic success through immersion and a part-time job, but they reflected that the DET had not adequately prepared them for the rigors of graduate-level English.

In retrospect, TT2 noted that if they were to repeat the process, they would opt for TOEFL, stating: *“Preparing for TOEFL prepares you more for a scholarly way of writing and comprehension, because the texts are more complex...if you prepare with time, I think the TOEFL is gonna give you more preparation than [the DET].”*

Discussion & Conclusion

For non-native English-speaking students intending to pursue higher education in English-speaking countries, proficiency test scores remain a critical prerequisite. Individual institutions establish minimum score requirements and determine which assessments are acceptable for admission. At the University of Minnesota and similar institutions, these benchmarks are set through a collaborative, committee-based process designed to ensure international students can thrive in a rigorous academic environment ([Lam et al., 2021](#)). While this is certainly a clear objective, ensuring student success involves complexities that extend beyond simply establishing a minimum score.

Educational researchers have long sought to develop language proficiency assessments that can accurately predict a test-taker's future academic performance. Predictive validity research has attempted to quantitatively demonstrate a connection between proficiency scores and academic success, typically measured by GPA. Initially, these analyses focused on 'gold standard' assessments such as the IELTS and TOEFL; more recently, DET has been included in these evaluations ([Isaacs et al., 2023](#)).

However, as demonstrated in the present study's literature review, there is no definitive evidence of a clear, universal connection between proficiency scores and academic achievement, regardless of the test or the era of research. That does not suggest a total absence of evidence; several studies reviewed here indicate a positive correlation and statistical significance for TOEFL, IELTS, and DET regarding GPA ([Wongtrirat 2010](#); [Dang & Dang, 2021](#); [Tabassum, 2023](#)). Nevertheless, sample limitations and inconsistent results across multiple variables preclude broad generalizations regarding predictive validity.

Researchers increasingly acknowledged that proficiency scores should not be viewed in isolation. While numerous factors contribute to an international student's ability to thrive, these scores remain an inescapable component of the university admissions process. As 'high-stakes' assessments, these tests carry direct consequences for the test-taker ([Wikipedia Contributors, 2018](#)). Consequently, the pressure to obtain a specific score is immense, often leading students to employ intensive preparation strategies. [Ma and Chong \(2022\)](#) found that Chinese students often rely on short-term preparation courses and 'leaked' test information to boost their results. While these techniques—often involving rote memorization and prepared answers—may improve scores, they may fail to develop authentic language proficiency. This disparity between the test scores and actual ability casts doubt on the test score's accuracy and, by extension, the student's readiness for academic success.

The present study seeks to investigate how English proficiency scores and students' perceptions of score accuracy influence their language confidence and future success.

Specifically, this paper addresses the following research question:

- **RQ1: What relationships exist between a test-taker's objective proficiency scores and their perceived language confidence regarding their readiness for collegiate academic study?**

Analysis of test-takers' responses indicates a positive relationship between CEFR scores and self-reported confidence. This relationship underscores the critical nature of *test validity*; specifically, test-takers base their self-efficacy on the assumption that the assessment provides an accurate measure of their proficiency.

If the test lacks validity—failing to measure the constructs it intends to—the resulting score becomes a poor proxy for actual ability. For instance, [Dang and Dang \(2021\)](#), found that certain IELTS writing and speaking tasks failed to replicate the specific linguistic demands of university-level assignments. In such cases, a high score may instill a false sense of confidence. This creates a “proficiency gap” where the student's perceived ability exceeds their actual performance, potentially leading to academic struggle when faced with rigorous college coursework. This disconnect is evident in TT2's experience with the DET: despite an academically ‘adequate’ score, their initial communicative struggles suggest the test may not have fully captured their functional language needs.

Confidence also varies significantly across specific language domains. In this study, nearly two-thirds (63.6%) of participants reported a lack of confidence in speaking, a figure that aligns closely with the 69.7% who scored in the CEFR B1-B2 range for that skill. By contrast, ‘not confident’ ratings for other skills were notably lower:

- Reading: 18.5%
- Listening: 30.3%
- Writing: 45.5%

This disparity indicates that many test-takers who feel secure in their receptive or written skills still struggle with oral production. While researchers like [Horwitz et al., \(1986\)](#) and [Krashen \(1982\)](#) have long identified speaking as the most difficult challenging skill for non-native speakers, its role remains paramount for both academic success and daily life in the United States.

The discrepancy raises questions about whether current proficiency tests adequately assess the specific speaking skills required for higher education. Interestingly, institutional requirements vary on this point; for example, the University of Minnesota requires minimum overall and writing scores but does not set a specific threshold for speaking, reading, or listening ([University of Minnesota, n.d.](#)). Further research is necessary to evaluate which assessment models most effectively measure speaking proficiency and predict academic readiness.

The second research question (*RQ2*) investigates the connections between test-takers' proficiency test scores and their perceptions of test accuracy regarding their preparation for collegiate study

While the 'Accuracy Feeling' reported by participants was generally high—with 94.3% of responses scoring a 3 or above—there is a notable relationship between higher CEFR scores and increased perceptions of accuracy. This phenomenon may be driven by psychological factors; research suggests that test-takers are more likely to validate the accuracy of an assessment when they achieve a desirable result, attributing the success to their inherent ability.

Conversely, when faced with lower scores, individuals often perceive the test as less accurate, frequently attributing the outcome to external factors or systemic flaws ([Zuckerman, 1979](#); [Okamoto et al., 2025](#)).

However, these perceptions of accuracy are often decoupled from objective academic measures. A student's GPA, for example, may not fully capture the subjective 'linguistic burden' they experience in the classroom. This disconnect was articulated by TT2, who initially viewed their proficiency test experience positively due to their high performance. Upon entering the university, however, they encountered a stark reality: the score did not translate to ease of use. TT2 reported having to exert disproportionate effort to manage academic readings and writing assignments, while simultaneously struggling to communicate effectively with faculty and peers.

While remedial language courses are an option, they are often rendered inaccessible by rigorous academic schedules or financial constraints. Furthermore, while immersion typically strengthens language skills over time, the immediate challenge remains significant. These findings place a heavy responsibility on educational assessment researchers and test developers. It is imperative to ensure that high-stakes English proficiency tests—taken by thousands of students annually—accurately reflect the functional demands of university life, rather than merely providing a score that satisfies an admissions requirement.

The final research question (RQ3) addresses the comparative perceptions of test-takers across the three primary assessments: IELTS, DET, and TOEFL.

Comparative Test Perceptions and Concurrent Validity

DET: Perceptions and Initial Validity

While the DET sample size is limited, initial data suggests that test-taker perceptions regarding confidence and accuracy align closely with those of the IELTS and TOEFL. For instance, DET confidence levels mirror the hierarchical patterns seen in the more established exams: participants reported the highest confidence in *Reading*, followed by *Listening*, *Writing* and *Speaking*, respectively. This alignment suggests a degree of *concurrent validity* between the DET and its counterparts. Furthermore, participants' belief in the accuracy of DET (66.7% rating it 4 or above), relates to the perceived accuracy of IELTS (70%) and TOEFL (60%).

Comparative Analysis: TOEFL vs. IELTS

The larger dataset for TOEFL and IELTS allows for a more granular comparison. When observing incoming *CEFR scores*, the distribution between the two tests is remarkably similar;. The modal scores for both overall scores and sub-skill categories were identical. For example, the highest percentage of overall CEFR score for both IELTS (60%), and TOEFL (65%) fell within the *C1 range*.

However, a distinct discrepancy emerged: TOEFL scores were consistently higher. When aggregating the 'high-tier' scores (C1 and C2), TOEFL's totals were at least 10% higher across all categories. This is largely attributable to the higher frequency of C2 results among TOEFL takers compared to IELTS takers (22 versus 2, respectively).

Concordance vs. Test Design

The discrepancy in high-tier scores raises a critical question: is the TOEFL inherently more accessible than the IELTS? Statistically, the two tests maintain a high correlation. A 2010 ETS concordance study of 1,153 participants found that one test is a reliable predictor of performance on the other, yielding a *Pearson's correlation coefficient (r)* above 0.8. This

finding was reaffirmed in a 2025 collaborative study by ETS and the British Council ([Ikeda et al., 2025](#)).

Despite this statistical correlation, the *construct and delivery* of these exams differ significantly:

- *TOEFL*: Favors students proficient in digital communication and those familiar with American accents.
- *IELTS*: Better suits students who prefer human interaction and possess high reading stamina

Such pedagogical differences suggest that individual personality and skill sets can lead to significant score differentials, potentially complicating the concept of test concordance.

Qualitative Insights: TT1 and TT2

The educational backgrounds of the interviewees (TT1 and TT2) reflect a broader trend in the dataset; 94.3% of surveyed students had studied English for over five years before arriving in the U.S. Despite this shared foundation, their experiences with the DET diverged:

Table 21

TT1 vs. TT2 Test Experiences

Metric	Test-Taker 1 (TT1)	Test-Taker 2 (TT2)
Comparative Tests	IELTS & DET`	TOEFL & DET
CEFR Range	B2	C1
DET Scores	Lower (relative to TT2)	10+ points higher than TT1
Self-Confidence	High in all areas	Lacking in Speaking/Listening
Perceived Accuracy	Mixed (Reading/Writing ‘inaccurate’)	High (All subscores ‘very accurate’)

Limitations

Sample Demographic and Generalizability

As an independent study, this research faces several constraints that impact the generalizability of the findings. A primary limitation is the *sample size*, a challenge also noted in recent literature ([Tabassum, 2023](#); [Isbell & McGehee, 2025](#)). Because recruitment was centered at a single university in Minnesota, the participant pool is relatively homogenous.

The sample is skewed toward *Social Science and Humanities (SSH)* majors relative to *Life, Physical Science and Engineering (LPSE)* majors. This limited variance may influence the results, as SSH coursework is typically more linguistically intensive, which could relate to higher linguistic confidence or more stringent self-assessments of accuracy. Additionally, because the vast majority of participants were graduate students, the sample reflects individuals who have had extensive opportunities to develop their English skills prior to entering U.S. higher education—a trend illustrated by the interviews with TT1 and TT2. Consequently, findings for incoming undergraduate students might differ significantly regarding linguistic confidence and perceptions of test accuracy. Finally, because this study could not bifurcate entry scores or perceptions by academic level, exploring these demographic-specific variations remains a critical avenue for future research..

Survey Design and Temporal Factors

The survey instrument did not account for *recency of testing*, which introduces potential *recall bias*. For example, interviewees TT1 and TT2 took the DET in 2021 and 2022,. By the time of their interviews, three to four years of residency in the U.S. had likely altered their perceptions of their initial proficiency and test experience.

Recommendation: Future studies should recruit participants who have completed their

proficiency exams within the last 6-12 months. This ensures the experience remains ‘fresh’ and minimizes the confounding variable of linguistic gains made during long-term immersion in the U.S.

Furthermore, the survey lacked a mechanism to track *recruitment sources*. While the majority of respondents likely originated from the researcher’s department, the anonymous nature of the data prevents a clear understanding of whether participants were reached via social media, department listservs, or personal networks. Identifying the source of respondents in future iterations would allow for a more nuanced demographic breakdown.

Disproportionate Sample Representation

A further limitation is the uneven sample distribution across the testing platforms. While the sample composition (TOEFL: 60.6%, IELTS: 30.3%, DET: 9.1%) roughly mirrors local application trends, it complicates a direct ‘equal footing’ comparison among the exams. Additionally, the qualitative interviews focused exclusively on DET test-takers rather than providing a comparative analysis of all three proficiency assessments.

To mitigate these limitations in future research, the following strategies are recommend:

- *Targeted Recruitment*: Researchers should explicitly recruit test-takers from underrepresented testing platforms (such as the DET) to achieve greater parity and statistical power.
- *Institutional Partnerships*: Collaborating with test-preparation organizations or specific English Language Programs (ELPs) could yield a more balanced and representative cohort.

Summary of Future Directions

While these limitations restrict the immediate application of the results to the broader international student population, they provide a roadmap for more robust inquiry. Expanding the study to multiple universities and implementing stricter controls on the testing recency will allow for a deeper understanding of how test-taker identity—be it through their major, degree level, or choice of exam—shapes their journey into U.S. higher education.

Based on the findings of the present study, several key questions warrant attention in future research. First, does a test-taker's linguistic confidence in linguistic ability influence their subsequent academic success? Second, what role do language background and experience play in English proficiency test performance and academic outcomes, and how can these factors be precisely measured? Finally, what are the most effective methods for preparing students—particularly in terms of speaking ability—for the rigors of higher education?

References

AmazingTalker. (2026). *How much does the IELTS test cost in 2026?*

https://en.amazingtalker.com/blog/en/other/122220/#How_Much_Does_the_IELTS_Test_Cost_in_2026

Andrade, M. S. (2006). International students in English-speaking universities: Adjustment factors and impacts on success. *Journal of Research in International Education*, 5(2), 131–154.

Arno. (2026). *How many people take the Duolingo English test?*

<https://goarno.io/blog/how-many-people-take-the-duolingo-english-test/>

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Barkaoui, K. (2025). The relationship between English language proficiency test scores and academic achievement: A longitudinal study of two tests. *Language Testing*, 42(3), 253-282. <https://journals.sagepub.com/doi/full/10.1177/02655322251319284>

Boddy, N. (2001). The revision of the IELTS Speaking test. *JALT Testing & Evaluation SIG Newsletter*, 5(2), 2–4.

British Council. (2020, April 24). IELTS has launched IELTS Indicator. Opportunities and Insight. <https://opportunities-insight.britishcouncil.org/short-articles/news/ielts-has-launched-ielts-indicator>

British Council. (2024). *What is IELTS?* <https://takeielts.britishcouncil.org/take-ielts/what-ielts>

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.

Bruce, E. (2023). IELTS Online: Overview. IELTS Research Group, 1–12.

<https://ielts.org/cdn/Research/ielts-online-overview-bruce-et-al-2023.pdf>

Burstein, J., Laffair, G., & von Davier, A. (n.d.). *A theoretical assessment ecosystem for a digital-first assessment: The Duolingo English test*. Duolingo.

<https://duolingo-testcenter.s3.amazonaws.com/media/resources/A+Theoretical+Assessment+Ecosystem+for+a+Digital-First+Assessment%E2%80%94The+Duolingo+English+Test.pdf>

Cambridge English. (2025). *Quality and accountability*.

<https://www.cambridgeenglish.org/english-research-group/quality-and-accountability/est.pdf>

Cambridge University Press & Assessment. (n.d.). *International language standards: The Common European Framework of Reference for Languages (CEFR)*.

<https://www.cambridgeenglish.org/exams-and-tests/cefr/>

Cambridge University Press & Assessment. (2017). IELTS data and historical growth milestones (1989–2015). International English Language Testing System.

<https://www.ielts.org/teaching-and-research/historical-data>

Canale, M., & Swain, M. (1980). Theoretical bases for communicative approaches to language teaching and testing. *Applied Linguistics*, 1(1), 1–47.

<https://doi.org/10.1093/applin/1.1.1>

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.

Cohen, M., & Teiser, S. (2025). *Living in the Chinese Cosmos*. Asia for Educators.

<https://afe.easia.columbia.edu/cosmos/irc/classics.htm>

Cooper, S. (2023). *Duolingo vs. IELTS: How to Decide*. Arno.

<https://goarno.io/blog/duolingo-vs-ielts-how-to-choose/>

Dabaliz, A. A., Kaadan, S., Dabbagh, M. M., Barakat, A., Shareef, M. A., Al-Tannir, M., Obeidat, A., & Mohamed, A. (2017). Predictive validity of pre-admission assessments on medical student performance. *International Journal of Medical Education*, 8, 408–413.

<https://doi.org/10.5116/ijme.5a10.04e1>

Dang, C. N., & Dang, T. N. Y. (2021). The predictive validity of the IELTS test and contribution of IELTS preparation courses to international students' subsequent academic study: Insights from Vietnamese international students in the UK. *RELC Journal*, 54(1), 84–98.

<https://doi.org/10.1177/0033688220985533>

Dawadi, S. (2020). Impacts of high-stakes English tests on students and parents: A washback study. *Language Testing in Asia*, 10(1), 1–21.

Duolingo. (2024). *Official guide for test takers*.

<https://englishtest-static.duolingo.com/media/resources/test-taker-guide/DET%20Official%20Guide%202024%20English.pdf>

Duolingo English Test. (2023, September 18). In *Wikipedia*.

https://en.wikipedia.org/wiki/Duolingo_English_Test

Educational Testing Service. (n.d.-a). *Getting your TOEFL iBT® scores*.

<https://www.ets.org/toefl/test-takers/ibt/scores/get-scores.html>

Educational Testing Service. (n.d.-b). *Sending your TOEFL iBT® scores*.

<https://www.ets.org/toefl/test-takers/ibt/scores/send-scores.html>

Educational Testing Service. (n.d.-c). *Why choose the TOEFL iBT test?*

<https://www.ets.org/toefl/test-takers/ibt/about/why.html>

Educational Testing Service. (1982). *The TOEFL native country file: Detailed data on candidate populations, 1977–1979* (TOEFL Research Report No. 11). Educational Testing Service.

Educational Testing Service. (2010). *Linking TOEFL iBT scores to IELTS scores*.

ETS.

https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf

Educational Testing Service. (2020). *The official guide to the TOEFL iBT test* (6th ed.).

McGraw-Hill Education.

Educational Testing Service. (2023). TOEFL program history (TOEFL Research Insight Series, Vol. 6). ETS Research Report Series.

<https://www.pt.ets.org/pdfs/toefl/toefl-ibt-insight-s1v6.pdf>

Farhady, H. (2018). History of language testing and assessment. In A. Lyon (Ed.), *The TESOL encyclopedia of English language teaching*. (pp. 1–7).

Fulcher, G. (2000). The “communicative” legacy in language testing. *System*, 28(4), 483–497.

[https://doi.org/10.1016/s0346-251x\(00\)00033-6](https://doi.org/10.1016/s0346-251x(00)00033-6)

Giri, A. (2010). *Language testing: Then and now*. ResearchGate.

https://www.researchgate.net/publication/228390033_Language_Testing_Then_and_Now

[W](#)

Graham, J. G. (1987). English Language Proficiency and the Prediction of Academic Success.

TESOL Quarterly, 21(3), 505. <https://doi.org/10.2307/3586500>

Gunadi. (2012). TOEFL in a brief historical overview from PBT to iBT. *Universitas Negeri*

Malang. <https://sastra.um.ac.id/wp-content/uploads/2012/01/1-Gunadi.docx.pdf>

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). *Summaries of studies involving the Test of English as a Foreign Language, 1963–1982* (TOEFL Research Report No. 16).

Educational Testing Service / College Entrance Examination Board.

- Hashemi, A., & Daneshfar, S. (2018). A review of the IELTS test: Focus on validity, reliability, and washback. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 3(1), 39–52. <https://doi.org/10.21093/ijeltal.v3i1.123>
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2), 125–132.
<https://doi.org/10.1111/j.1540-4781.1986.tb05256.x>
- IELTS. (2024). *Trusted around the world—IELTS continues to lead, support, and empower*.
<https://ielts.org/news-and-insights/ielts-trusted-by-millions-in-2023>
- Ihlenfeldt, S. D., & Rios, J. A. (2022). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, 40(2), 026553222211123. <https://doi.org/10.1177/02655322221112364>
- Ikeda, N., Clark, T., Papageorgiou, S., Gu, L., Ohta, R., Blackhurst, A., & Bruce, E. (2025). *Aligning scores of language proficiency tests: A score concordance study between IELTS Academic and TOEFL iBT*. IELTS.
<https://ielts.org/researchers/our-research/research-reports/score-concordance-study-between-ielts-academic-and-toefl-ibt>
- International English Language Testing System. (2017). In Wikipedia.
https://en.wikipedia.org/wiki/International_English_Language_Testing_System

- Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university. *Language Testing*, 40(3), 026553222311585. <https://doi.org/10.1177/02655322231158550>
- Isbell, D. & McGehee, M. (2025). Relationships between English proficiency test scores and academic outcomes: DET, IELTS, and TOEFL in a U.S. public research-intensive university. *Studies in Language Assessment*, 14(1).
https://arts.unimelb.edu.au/_data/assets/pdf_file/0004/5347822/2025_McGehee_Isbell_Early-View.pdf?ref=blog.englishtest.duolingo.com
- Issuu. (2022). *Adopting the Duolingo English Test: Two institutions share their challenges and successes*.
https://issuu.com/nagapgem/docs/perspectives_34_1_spring_22/s/14973123
- Knoch, U., & Fan, J. (2024). The validity of score concordance tables: Principles and practices. *Language Testing*.
<https://journals.sagepub.com/doi/10.1177/02655322241239348>
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.
https://www.sdkrashen.com/content/books/principles_and_practice.pdf
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. McGraw-Hill.
- Laerd Dissertation. (2012). *Purposive Sampling*.
<https://dissertation.laerd.com/purposive-sampling.php>

Lam, D. M. K., Green, A., Murray, N., and Gayton, A. (2021). *How IELTS scores are set and used for university admissions selection*. IELTS Research Reports Online Series 2021/3. <https://ielts.org/cdn/Research/how-ielts-scores-set-and-used-for-university-admissions-selection-lam-et-al-report-2021.pdf>

Lardinois, F. (2014, July 23). *Duolingo launches its certification program to take on TOEFL*. TechCrunch. <https://techcrunch.com/2014/07/23/duolingo-launches-its-language-certification-program/>

Language Testing International. (n.d.). *Understanding language proficiency*. <https://www.languagetesting.com/lti-information/understanding-proficiency>

Lumetest. (2025). *Duolingo English Test (DET) vs IELTS: Which One To Choose?* <https://lumetest.com/blog/duolingo-english-test/duolingo-english-test-vs-ielts>

Ma, H., & Chong, S. W. (2022). Predictability of IELTS in a high-stakes context: A mixed methods study of Chinese students' perspectives on test preparation. *Language Testing in Asia*, 12(1), Article 3. <https://doi.org/10.1186/s40468-021-00152-3>

Manhattan Review. (n.d.). *IELTS History*. <https://www.manhattanreview.com/ielts-history/>

Mao, A. (2022). Literature review of language testing theories and approaches. *Open Access Library Journal*, 9, 1–5. <https://doi.org/10.4236/oalib.1108741>

Market Reports World. (2024). IELTS practice and exam platform market size & industry report, 2033 (Report ID: 14720232).

<https://www.marketreportsworld.com/market-reports/ielts-practice-and-exam-platform-market-14720232>

McNamara, T. (2011). Measuring second language performance. *Language Testing*, 28(4), 500–515.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–157). Oxford University Press.

Naismith, B., Cardwell, R., Laflair, G., Nydick, S., & Kostromitina, M. (n.d.). *Duolingo English test: Technical manual*.

https://duolingo-papers.s3.us-east-1.amazonaws.com/other/technical_manual/DET_technical_manual_2025_07.pdf

Okamoto, N., Taylor, M., Kubo, T., Ishii, S., De Martino, B., & Cortese, A. (2025). Blaming luck, claiming skill: Self-attribution bias in error assignment. *PLOS Computational Biology*, 21(12), e1013787. <https://doi.org/10.1371/journal.pcbi.1013787>

Oller, J. W. (1979). The Psychology of Language and Contrastive Linguistics: The Research and the Debate. *Foreign Language Annals*, 12(4), 299–309.

<https://doi.org/10.1111/j.1944-9720.1979.tb00188.x>

Prodigy Finance. (2023). *What you need to know about the Duolingo Language Test*.

<https://prodigyfinance.com/resources/blog/duolingo-language-test/>

Ramanathan, T. (2025, May 13). Duolingo | App, Languages, Background & Virality.

Encyclopedia Britannica. <https://www.britannica.com/topic/Duolingo>

Ren, H., et al. (2026). *Comparing new TOEFL 2026 with former TOEFL 2023 and IELTS*.

IELTS.

<https://ielts.org/researchers/our-research/research-reports/comparing-new-toefl-2026-with-former-toefl-2023-and-ielts>

Roemer, A. (2002). A more valid alternative to TOEFL? *College and University*, 77(4), 13–17.

Sadeghi, K., Ballıdağ, A., & Mede, E. (2021). The washback effect of TOEFL iBT and a local English Proficiency Exam on students' motivation, autonomy and language learning strategies. *Heliyon*, 7(10), e08135. <https://doi.org/10.1016/j.heliyon.2021.e08135>

Shea, L., Pesa, J., Geonnotti, G., Powell, V., Kahn, C., & Peters, W. (2022). Improving diversity in study participation: Patient perspectives on barriers, racial differences, and the role of communities. *Health expectations: An international journal of public participation in health care and health policy*, 25(4), 1979–1987.

<https://doi.org/10.1111/hex.13554>

Study.com. (2026, April). *The cost of TOEFL in 2026: All fees explained*.

<https://study.com/toefl/registration/how-much-does-the-toefl-cost.html>

Svensson, P. (2025, November 28). Keeping the Streak Alive: The Story of Duolingo.

Quartr.com. <https://quartr.com/insights/edge/keeping-the-streak-alive-the-story-of-duolingo>
[o](#)

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford University Press.

Spolsky, B. (2008). Language assessment in historical and future perspective. In N. H. Hornberger (Ed.), *Encyclopedia of language and education*. Springer,
https://doi.org/10.1007/978-0-387-30424-3_192

Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, 42(1), 1–40.
<https://doi.org/10.1017/s0261444808005399>

Sudarsono, S. (2021). The historical trajectory of TOEFL structural overhauls and its washback on international candidate strategies. *Indonesian Journal of Applied Linguistics*, 11(2), 412–425.

Tabassum, A. (2023, June). Pandemic Precipitousness: The Predictive Validity of Duolingo English Test for Student Success. Scholaris.ca.
<https://ucalgary.scholaris.ca/items/d73962ed-4985-4144-9111-c76f3d536d5c>

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. *ETS Research Report Series*, 2008(2), i–75.

<https://doi.org/10.1002/j.2333-8504.2008.tb02120.x>

Tomedes. (n.d.). *Language proficiency test*.

<https://www.tomedes.com/translator-hub/language-proficiency-test>

University of Minnesota. (n.d.). *English proficiency information*.

<https://admissions.tc.umn.edu/admissions/international-admission/english-proficiency-information>

University of Minnesota. (2025). *Count of U of M English proficiency submissions by test type and academic year, 2020/21 - 2024/25*. [Unpublished raw data].

Weir, C. J., O'Sullivan, B., & Yin, J. (2007). A comparative study of the cognitive processes of test takers in paper-based and computer-based IELTS reading. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment* (pp. 416–454). Cambridge University Press.

Wikipedia Contributors. (2018, November 15). *High-stakes testing*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/High-stakes_testing

Wongtrirat, R. (2010). *English language proficiency and academic achievement of international*

students: A meta-analysis. College and University, 86(2), 39–48.

<https://doi.org/10.25777/y7yt-m587>

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, 47(2), 245–287.

<https://doi.org/10.1111/j.1467-6494.1979.tb00202.x>

Appendix A

English Proficiency Test Survey

The following survey asks for information about your preparedness for studying in the U.S., your proficiency test scores, and your feelings about the accuracy of the test. Please select "Next" and complete Part 1 and 2. Then follow the directions in Part 3 to submit the survey. Thank you for participating!

PART 1: Preparation for Study

- How many years did you study English before you came to the U.S?
0-1 year / 1-3 years / 3-5 years / 5+ years
- What level of schooling did you start at in the U.S.?
High school / undergraduate / graduate
- Did you have to take English classes in the U.S. before starting your schooling?
Yes / no
- In which areas do you feel your English ability was inadequate for study in the U.S. when you started? (choose as many as apply)
Reading / writing / speaking / listening / my ability was adequate
- What is your current major of study?

PART 2: Proficiency Test Scores and Feelings

• Which proficiency test did you use to come to the U.S. for study?
International English Language Testing Standard (IELTS) / Test of English as a Foreign Language (TOEFL) / Duolingo English Test (DET) / Other:

- What was your overall test score?

- To what extent do you feel your proficiency test score is an accurate reflection of your English ability?

Not at all 1 2 3 4 5 Very much so

- Why do you think the test score was or was not accurate?
- _____
What was your reading subscore?
- _____
What was your writing subscore?
- _____
What was your speaking subscore?
- _____
What was your listening subscore?

- How accurately do you think each of the subscores reflected your preparedness for study in the U.S.? (choose one accuracy point for each subscore)

	not accurate	somewhat accurate	very accurate
Reading			
Writing			
speaking			
listening			

- Do you feel preparing for and taking the proficiency test helped prepare you for study in the U.S.?

Not at all 1 2 3 4 5 Very much so

- Why do you think preparing for and taking proficiency test did or did not help prepare you for study in the U.S.?
- What do you feel has helped you improve your English the most (before or after coming to the U.S.)?

Part 3

The author of this survey wishes to interview two survey participants about this topic. They will participate in a 30 minute Zoom interview about their testing and U.S. educational experience. If you are willing to be considered as an interview participant, please choose "Yes" below. On the next page you will be asked to type your contact email. Only those chosen to participate in the interview will be contacted. If you do not wish to be considered, please choose "No" below. You will be taken directly to the submission page.

- Are you willing to participate in an interview on this topic?
Yes / no

Contact Info for Interview

Thank you for your willingness to participate in a follow-up interview.

- Please type your contact email address.

Appendix B

Interview Questions

- What is your current education level?
- What is your major?
- How long have you been in the U.S.?
- When did you take the English proficiency test?
- Describe your English learning experience before coming to the U.S.
- Did you do any preparation specifically for the proficiency test? Describe it.
- Why did you choose that particular proficiency test?
- Share your experience of taking the proficiency test. How did you feel at that time?
- After taking the proficiency test, did you feel confident in your English ability? Why or why not?
- Did you take any English classes or do any language preparation after coming to the U.S. before starting your university classes? Describe it.
- Describe your first few days of university classes. Did you feel your English ability was adequate? How did you cope with the language difficulties?
- Have your feelings changed about English and your language ability since the beginning? How so? What do you think has brought about the changes?
- Do you think the proficiency test content correlates well with what you are learning in university? Why or why not?
- If you could start over again, how would you change your language preparation? Would you take a different proficiency test? Why?

Appendix C

Data Point Combinations for Analysis

1. Total CEFR Scores—how many of each score (B1, B2, C1, C2) there were for overall proficiency scores and for each subscore
2. Confidence Level—for each subscore, how many were “confident,” and how many were “not confident”?
3. CEFR Scores & Confidence Level—For each subscore and each CEFR score, how many were “confident” and “not confident”?
4. CEFR Scores By Test—How many of each CEFR score (overall score and subscores) there were, separated by proficiency test (DET, IELTS, TOEFL)
5. Confidence Level By Test—For each subscore, how many were “confident” and “not confident,” separated by proficiency test
6. CEFR Scores & Confidence Level By Test—For each subscore and each CEFR score, how many were “confident” and “not confident,” separated by proficiency test
7. CEFR Scores By Major Category—how many of each CEFR score (overall and subscores) there were separated by major category (SSH, PSE)
8. Confidence Level By Major Category—for each subscore, how many were “confident” and “not confident,” separated by major category
9. CEFR Scores & Confidence Level By Major Category—for each subscore and each CEFR score, how many were “confident” and “not confident,” separated by major category
10. US English Class By CEFR Score—For each CEFR score, how many were required to take English before starting their major, and how many were not
11. US English Class By Major Category—For each major category, how many were required to take English before starting their major, and how many were not
12. Accuracy Feeling By CEFR Score—For each CEFR overall score, how accurate did respondents feel the test was (on a scale of 1 to 5)?
13. Accuracy Feeling By CEFR Subscore—For each CEFR subscore, how accurate did respondents feel the test was (not accurate, somewhat accurate, very accurate)?
14. Accuracy Feeling By Test—For each proficiency test, how accurate did respondents feel the test was (on a scale of 1 to 5)
15. Accuracy Feeling vs. Preparatory Ability—for each level of accuracy (1 to 5), how well did respondents believe the test helped prepare them for university study
16. CEFR Score vs. Preparatory Ability—For each CEFR score (B1 to C2), how well did respondents believe the test helped prepare them for university study
17. Test VS Preparatory Ability—For each proficiency test (DET, IELTS, TOEFL), how well did respondents believe the test helped prepare them for university study