MICHIGAN LANGUAGE ASSESSMENT

Concordance Report

Score Equivalences Between MET and IELTS Academic

February 2025

Michigan Language Assessment
Dr. Svetlana Cook
Dr. Mika Hoffman
Patrick McLain

Independent Consultant
Dr. Hanan Khalifa

Last edited March 23, 2026

Table of Contents

1.0 Introduction

Test scores on foreign language proficiency tests are widely used to determine language learners' ability. Although many tests report score equivalences to commonly used scales of proficiency such as the Common European Framework of Reference (CEFR), it is often necessary to establish equivalences between scores on different tests. The following study investigates the score equivalence between the Michigan English Test (MET) and the International English Language Testing System (IELTS) Academic test.

Best practices for concordance studies include establishing that the two tests are suitable for linking, implementing an appropriate methodology, and publishing both technical information and information targeted at test score users (Knoch & Fan, 2024). The Appendix lists good practice principles and how this study has implemented them.

2.0 Rationale for the Selection of IELTS Academic

General best practice in concordancing, following Kolen & Brennan (2014) and Knoch & Fan (2024), is to establish that the two tests are sufficiently similar in inferences, constructs, populations, and measurement characteristics. MET and IELTS Academic have long been used for similar purposes and as such would be expected to have a substantial degree of similarity. Although IELTS is more widely used, MET scores, like IELTS scores, are accepted by a wide range of government agencies, academic institutions, licensing boards, and other recognizing organizations. Both tests were developed with the involvement of what is now Cambridge University Press & Assessment, taking their theoretical basis from the same roots as the CEFR (Council of Europe, 2001, 2020). The tests are both multi-level exams that assess the same four skills at approximately the same ranges of the CEFR and report scores for the four skills separately as well as reporting an overall score.

2.1 Inferences

Both MET and IELTS Academic are tests of English language proficiency intended for high-stakes uses such as immigration and university admissions; although IELTS has a General Training version used for nonacademic purposes, the IELTS Academic is commonly used both for academic and professional purposes. Each test reports separate scores for listening, reading, speaking, and writing, as well as an overall score that is an average of the four sections. IELTS reports band scores on a scale that has a formal conversion to the CEFR (IELTS, 2024a; Lim et al., 2013); MET reports scores directly using the CEFR, also based on formal standard-setting (Michigan Language Assessment, 2014; Papageorgiou, 2010). The alignment of both tests to the CEFR and the common test purposes indicate that the inferences intended to be drawn from the two tests are essentially the same.

2.2 Construct

Overall, both tests address language skills important for success in academic and professional life, such as the ability to understand main ideas and details and the ability to state and defend opinions. Both MET and IELTS measure language proficiency broken down into four sections corresponding to the traditional language skills of listening, reading, speaking, and writing. Both tests are designed to assess communicative competence as set forth in the descriptor scales of the CEFR and in Bachman & Palmer (1996). This section compares how the skills are operationalized in the tasks for each test.

2.2.1 Listening

The MET Listening Section measures the ability to understand main ideas, details, specific information, and implications in a variety of contexts. It consists of three parts: discrete items based on short dialogues, sets of items based on longer dialogues, and sets of items based on a monologue. Topics cover personal, public, occupational, and educational contexts, and tasks focus on overall comprehension, understanding details, and understanding implied meaning. The stimuli are played once. All items are selected response, with the items presented on the screen. Items within sets are presented in a fixed order according to the appearance of the concept in the stimulus; discrete items are presented in random order.

The IELTS Listening Section measures the ability to understand main ideas, details, specific information, directions, and implications in a variety of contexts. It consists of four parts: two with sets of items based on dialogues, and two with sets of items based on monologues. Topics cover everyday situations and professional and academic contexts, and tasks focus on overall comprehension, understanding details, and relating parts of the stimulus. The stimuli are played once. There are a variety of item types, all of which are presented on the screen or in the test booklet; some require writing short answers.

The two tests use different formats for eliciting responses, but the overall format of having test takers listen to a monologue or dialogue and then checking for comprehension of global and detailed information is similar.

2.2.2 Reading

The MET Reading Section measures knowledge of grammatical features and the ability to understand main ideas, details, and implications, and to relate different texts or parts of texts to each other. It consists of three parts: discrete items focused on supplying correct vocabulary and grammar, sets of items based on extended reading passages, and sets of items based on three related reading passages. Topics cover personal, public, occupational, and educational contexts, and tasks focus on overall comprehension, understanding details, scanning, understanding grammar, and understanding implied meaning. All items are selected response.

The IELTS Reading Section measures the ability to understand main ideas, recognize specific information and relationships between parts of texts, and understand the difference between main ideas and supporting details. It consists of three sets of items based on extended reading

passages, with several different item types, including selected response, matching, and writing short answers. Tasks focus on overall comprehension, understanding details, scanning, and understanding opinions and implied meaning.

There are a few differences between how the tests represent the construct: MET includes items explicitly testing knowledge of grammar and vocabulary, which IELTS does not, and MET also includes items testing cross-text comparisons, which IELTS does not. IELTS explicitly tests the structure of longer texts with tasks putting paragraphs in order, which MET does not. Similarities between the MET Reading Section and IELTS Reading Section include presenting extended reading passages with a variety of items focusing on understanding main ideas, details, and opinions or implied information. Overall, there is considerable overlap between the two tests in the aspects of reading tested.

2.2.3 Speaking

The MET Speaking Section measures skills relating to several illustrative scales at multiple levels of the CEFR, including giving descriptions, narrating an experience, and expressing and supporting opinions. The section consists of three parts, with five tasks: the first part has three tasks, in which test takers see a picture on the screen and are asked to describe the picture, then narrate a personal experience related to the topic of the picture, then state and support an opinion or preference related to the topic of the picture. In the second part, the task is to discuss advantages and disadvantages of several options. In the third part, the task is to argue for a point of view or proposal. The speaking prompts are delivered by a recorded voice, and test takers speak into a microphone to provide their answers; their responses are later rated centrally by trained human raters according to rating criteria for task completion, language resources (grammar and vocabulary), and intelligibility/delivery.

The IELTS Speaking Section tests the ability to converse on everyday topics, organize a discourse, and support and explain opinions. It consists of three parts: an initial interchange between the test taker and an examiner in which the test taker asks general questions on common topics, the test taker is then given a written prompt to discuss a specific topic, and the examiner asks follow-up questions following the test taker's discussion. The test is rated in real time by the examiner, based on rating criteria for fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation.

The major differences between the tests are that IELTS involves interaction, with test takers responding to questions and engaging in a conversation, whereas MET does not, and MET's picture description task elicits the ability to describe objects and situations, as it is geared toward lower CEFR levels, whereas IELTS does not. Another difference is that MET's rating criteria include task completion, which those for IELTS do not. The two speaking constructs are similar in requiring a monologue for a substantial portion of the test, so that they elicit a sample long enough to assess organization and coherence, and they also both provide tasks that test the ability to speak on everyday, familiar topics, to organize information, and to support opinions. The rating criteria also overlap to a large degree, assessing grammar, vocabulary, and fluency. Overall, the two tests show considerable overlap in the aspects of speaking tested.

2.2.4 Writing

The MET Writing Section measures the ability to describe a personal experience, express and elaborate upon an opinion, and support a point of view with comparison and contrast. It consists of two parts. The first part requires test takers to provide three short related texts describing a personal experience on a given topic, a personal opinion on a related topic, and an elaboration on the opinion; these tasks target CEFR levels A2 and B1. The second part is an essay requiring support for a position, targeting CEFR levels B2 and C1. Responses are rated centrally by trained human raters according to rating criteria for task completion, grammatical accuracy, vocabulary, mechanics, and cohesion/organization.

The IELTS Writing Section measures the ability to describe and present facts and to support a point of view with comparison and contrast. It consists of two tasks: the first requires test takers to describe a visual prompt, and the second is an essay requiring support of a point of view. Responses are rated by trained raters according to rating criteria for task achievement/response, coherence/cohesion, lexical resource, and grammatical range and accuracy.

The main difference between the Writing sections of the two tests is that IELTS requires a response to a visual prompt, whereas MET does not.The two writing constructs are generally very similar: the first task elicits factual information and is geared towards lower CEFR levels, and the second is an essay requiring support for a position. The rating criteria are also very similar, although the IELTS scoring rules weight the essay more heavily than the description task, whereas MET scoring rules weight both parts the same.

2.3 Population

Both MET and IELTS Academic are taken for multiple purposes, including academic admission, professional certification, and immigration. Both tests are taken primarily by adults and high school leavers from a variety of language backgrounds and are administered in a large number of countries around the world. For the purposes of this study, the populations are suitably similar in terms of demographics and educational and linguistic background.

2.4 Measurement Characteristics

Measurement characteristics are essentially the ways in which the test constructs are operationalized, for example, test length, administration conditions, and task type (Kolen & Brennan, 2014). Both MET and IELTS are high-stakes tests administered in secure, proctored testing environments in authorized and regularly inspected test centers. (Note that both tests provide the option of testing online with remote proctoring, but as these test versions are not always accepted for high-stakes uses, no participants in this study took remote-proctored versions.) IELTS also has a paper version, although virtually all participants taking IELTS for this study took the digital version. Both tests are delivered in four separately timed sections, as summarized below:

Table 2.4.1

Test Summary

Section	MET		IELTS	
	Time	Number of Questions	Time	Number of Questions/Tasks
Listening	35 minutes	50	30 minutes	40
Reading	65 minutes	50	60 minutes	40
Writing	45 minutes	4	60 minutes	2
Speaking	10 minutes	5	11–14 minutes	3
Total	155 minutes	109	161–164 minutes	85

The timing and number of questions are comparable; as indicated in the construct section (2.2), the task types are somewhat different. MET listening and reading sections have selected-response questions only, for example, whereas IELTS has both constructed and selected response types. MET speaking is recorded and rated centrally after the exam, whereas IELTS speaking involves a live examiner. The writing sections are the most similar, with both tests having an essay requiring support for a position and a section focusing on shorter, more factual writing, though they differ in the type of writing in the shorter section: MET tasks require a description in response to a written prompt, whereas IELTS provides a visual prompt. Both MET and IELTS have robust quality control systems for their human marking, including double-marking a proportion of Writing and Speaking responses.

2.5 Familiarity and Test Preparation

The testing organizations responsible for MET and IELTS Academic are committed to helping examinees become well-prepared and confident before taking their exams by making a range of preparation resources available at no cost on their respective websites. For those taking MET, there is an array of resources designed to familiarize test takers with the exam's structure and content. These resources include sample tests that simulate the actual exam experience, and study guides that offer strategies and tips for effective preparation. Additionally, MET provides resource packs that break down the test into its four main sections, helping candidates understand what each part entails. Videos showcasing test takers' performances are available to provide a practical example of what is expected, while annotated writing samples at various proficiency levels serve as a benchmark for self-assessment. Practice test books are also available for purchase.

Similarly, IELTS offers its own comprehensive suite of preparation tools. These include official sample tests that mirror the real exam, allowing candidates to practice under similar conditions. For those who need to focus on writing, IELTS provides specific training videos tailored to this section of the test. Beyond these, IELTS also offers links to a broad spectrum of additional resources such

as online courses, interactive applications, and webinars, as well as a library of videos, books, and articles designed to deepen candidates' understanding and enhance their skills.

Both MET and IELTS prominently feature references to preparation materials on their respective test registration pages. This ensures that test takers are fully aware of the multitude of resources at their disposal. By explicitly highlighting preparation materials, the testing organizations aim to guide candidates in learning about the different sections of the exams, the variety of tasks and items included, the overall format, and the timing constraints they will encounter, ultimately aiding them in their journey toward successful test performance.

At the time of the concordance study, most participants were studying English or preparing for one of the exams; we worked closely with the English training providers to ensure that both teachers and their students were familiar with the MET exam and had access to preparation and practice materials, which included (1) an overview of the sections and types of items or tasks of each test, (2) sample tests to experience the content and format of the test, and (3) guidelines to prepare for the different sections of the tests. Access to these resources was also shared with individual participants after they registered for the study.

2.6 Reporting Scores

MET comprises four sections: Listening, Reading, Speaking, and Writing. The MET Listening and Reading Sections are machine scored through the test delivery platform. Each correct answer contributes to the final score for each section, and there are no points deducted for wrong answers. A scaled score, ranging from 0 to 80, is calculated using Item Response Theory. This method ensures that scores are comparable across different forms and administrations, and that the ability required to receive a score remains the same over time. The Writing and Speaking Sections are graded by human raters according to scales established by Michigan Language Assessment and published on the Michigan Language Assessment website. All raters are trained and certified by Michigan Language Assessment specifically for the particular skill and rating scale, and a percentage of responses are rated by two raters for quality assurance purposes.

MET uses the same scale for all four components of the test. All test takers receive a scaled score from 0–80 for each test section and an overall average score for all sections taken. The overall score is a truncated integer of the subtest average. The scores are also reported as CEFR levels. The table below shows the MET scaled scores that correspond to these CEFR levels. These correspondences are based on standard-setting research conducted by Michigan Language Assessment (Papageorgiou, 2010; Michigan Language Assessment, 2014).

Table 2.5.1

MET-CEFR Correspondence

CEFR Levels	MET Scaled Scores
C1 or above	64–80

CEFR Levels	MET Scaled Scores
B2	53–63
B1	40–52
A2	27–39
Bellow A2	0–26

IELTS results are reported on a nine-band scale (including half bands), ranging from *Non-user* of the language (Band 1) to *Expert* language user (Band 9). Each section is scored individually. The overall band score is the average of the four section band scores rounded to the nearest half band. IELTS is aligned to CEFR; however, the IELTS band scores do not align exactly with the CEFR transition points. As can be seen in the image below, an IELTS band score of 5 may be associated with either CEFR B1 or B2, and an IELTS band score of 6.5 may be associated with either CEFR B2 or C1. Both tests thus span most of the range of the CEFR, and thus report scores at similar ability levels.

Figure 2.5.1
IELTS-CEFR Correspondence

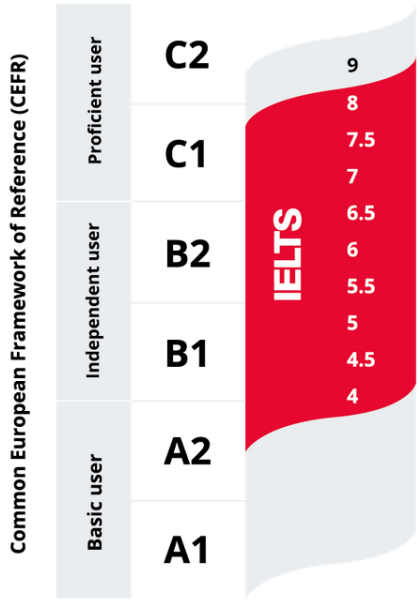


Image from IELTS (2024a), “IELTS and the CEFR,”
https://ielts.org/organisations/ielts-for-organisations/compare-ielts/ielts-and-the-cefr

2.7 Reliability

The reliability of the linked tests is a major factor in the accuracy and interpretation of the linking results. Reliability is important because it is an indication of how confident score users can be that the scores consistently represent the same level of ability. Good practice in concordancing is that both tests should have high reliability, and the reliabilities of the two tests should be roughly equal. MET and IELTS are high-stakes tests that have rigorous processes in place to assess the tests' reliability and validity. Based on publicly available information, the average reliability estimate for MET Listening is 0.88 with a standard error of measurement (SEM)[1]= 3.90 (MET 2023 Test Report). IELTS Listening has a reported reliability of 0.92[2] and SEM of 0.38 (IELTS, 2024). For MET Reading, the average reliability estimate is 0.86, and the SEM estimate was 4.67. The reliability of IELTS Reading (Academic) was 0.89 and the SEM estimate was 0.41. These reliabilities are close, indicating a similarity in measurement quality between the two tests.

For the productive skills, IELTS reports inter-rater reliability coefficient of 0.88 for Speaking and 0.95 for Writing based on the operational double-rating data for 2023–2024. IELTS does not report SEM for the productive skills. The quality of MET raters has been evaluated by the Intraclass Correlation Coefficient (ICC), which can be directly compared with the inter-rater reliability coefficient. The MET ICC coefficient is 0.90 for Speaking (SEM = 5.16), and 0.86 for Writing (SEM = 4.92). Both tests have high reliability, and the reliabilities of the two tests are similar, so, as with the Listening and Reading tests, MET and IELTS have similar measurement quality.

The overall SEM for MET was computed as an average of the subtest CSEMs across all MET score points. The test's overall SEM is 2.72. The overall SEM for IELTS is 0.17. When comparing SEMs of two tests, it is important to keep in mind that SEM is not a standardized unit of measure, and that its magnitude is directly tied to the size of the scale. All tests have some measurement error; SEM gives an indication of the amount of variation we might see in scores from a test taker who takes the same test multiple times. MET's 80-point scale is about five times larger than IELTS reporting scale (17-point scale, 1 to 9 in 0.5-point increments), which is reflected in the SEM. The SEM values of the two tests are comparable after accounting for the discrepancies in the size of the scale between the two tests, and, therefore, the tests are appropriate for linking.

3.0 Methodology

Consistent with recommendations in Knoch & Fan (2024), the study used a counterbalanced-group design and equipercentile linking with smoothing to empirically determine the minimum cut score linkage between IELTS and MET. Participants took both tests in person at

[1] Scoring of MET Listening and Reading employs IRT methodology, which utilizes CSEM (Conditional Standard Error of Measurement) instead of a traditional SEM, a statistic common within the Classical Test Theory. CSEM is different from SEM in that it varies along the continuum of the scale. Here we report an SEM as an averaged value of CSEMs calculated for each score point on the MET scale for each section.

[2] IELTS reliability measures for Listening and Reading are not directly available on the test's website; the estimates were back-transformed from the published statistics (SD and SEM). (Retrieved on 10/29/2024 from https://ielts.org/researchers/our-research/test-statistics#Test_performance)

a secure and authorized test center, and each participant took the two tests within three months of each other, with most participants taking both tests within one month.

3.1 Sample Population

Michigan Language Assessment collected official scores for IELTS and MET from 1,066 participants, who were recruited from partner organizations, authorized test centers, and education programs and completed an application form. The application form was used to obtain permission for their data to be used, to provide demographic and other information to ensure we were targeting a representative sample, and to ensure that if a participant had previously taken one of the tests, the test had been taken within three months of when they were scheduled to take the other test. They also provided demographic information and information about their purpose for taking the test. Some participants were offered a financial incentive to participate, as they might not have otherwise taken both tests.

Michigan Language Assessment and its partners worked with participants to promote a balance between participants who took IELTS first and those who took MET first. More participants took MET first than took IELTS first (58% versus 42%); data collection relied heavily on participants who had already completed testing for one test, and the one test was generally MET, given that the participants came from Michigan Language Assessment partners. We conducted an analysis of a subset (n = 722) of the 1,066 participants in which counterbalancing was exactly at 50 percent and found that the results were essentially the same as in the larger cleaned sample (n = 1,000; see section 3.3); the larger sample was used for this study to provide more robust score equivalences. We also conducted an invariance study on test order that showed that the equating results did not vary significantly depending on the order of taking the tests (see section 4.6.4), so the departure from a perfect counterbalancing did not affect the results.

Our partners leveraged an array of free preparation resources available on the MET and IELTS websites to help test takers gain a better understanding of the exams' content and format. By directing participants to these resources, they ensured that candidates could access comprehensive preparatory materials offered by the official testing organizations or affiliated learning providers. Moreover, many partner organizations that assisted in recruiting participants for the study provided additional educational and exam familiarization opportunities. They offered English and test preparation courses at their institutions, creating more avenues for participants to enhance their skills and feel prepared to take both exams. In such cases, participants were also encouraged to make use of the resources and services provided by our partner organizations. The impact of these efforts was noticeable in the increased accessibility and utilization of MET's digital sample tests, which are available free of charge on the Michigan Assessment website. Data indicated a significant rise in engagement with these resources during the period in which testing for this study took place, thereby demonstrating the success of the resource-sharing strategy.

Specific information about the sample population is covered in section 4.1.

3.2 Data Collection

All MET and IELTS test scores used in this study are trustworthy, as they were obtained from official score reports for tests that took place at accredited test centers. MET scores were processed directly by Michigan Language Assessment. Participants were required to submit official IELTS test report forms as a condition of participation in the study.

3.3 Data Cleaning

The data collection yielded a dataset of 1,066 test takers, distributed across the majority of the scale ranges for both tests. IELTS scores in the sample ranged from Band 3 to Band 9 (mean = 6.53, median = 6.5); MET scores ranged from 23 to 80 (mean = 57.86, median = 58). Descriptive statistics—score distributions and measures of central tendency (mean, Standard Deviation (SD), and Standard Error (SE)) by IELTS band—are reported in Table 3.3.1.

Table 3.3.1

Descriptive Statistics of MET Scores by IELTS Band, n = 1,066

IELTS band	n	mean	min	max	range	SD	SE
3	1	34.00	34	34	0	N/A	N/A
3.5	2	29.50	26	33	7	4.95	3.50
4	5	35.20	31	42	11	4.60	2.06
4.5	21	41.19	30	52	22	4.75	1.04
5	74	44.96	28	54	26	5.03	0.58
5.5	117	48.76	34	65	31	4.48	0.41
6	197	53.40	26	65	39	4.88	0.35
6.5	215	58.33	23	73	50	5.80	0.40
7	173	61.74	52	76	24	5.18	0.39
7.5	148	66.24	55	77	22	4.78	0.39
8	85	69.68	59	80	21	4.82	0.52
8.5	27	73.19	65	80	15	4.06	0.78
9	1	74	74	74	0	N/A	N/A

Out of concern that some of the test takers might not have had the same motivation in completing both tests—which might result in an inaccurate representation of the population and violate the equivalent group study design assumption—the data was analyzed for outliers. By removing outliers from the sample, Michigan Language Assessment attempted to correct for the discrepancy potentially introduced by the motivation variable.The outliers were operationalized in three main ways: discrepancy by overall score, discrepancy by specific subtest score, and by falling outside the range.

Discrepancy by overall score was operationalized as those test takers whose overall MET scores fell more than two SD of the mean overall MET score averaged across the IELTS half-band scores, corresponding to a 95% confidence interval.

Discrepancy by specific score was operationally defined as a set of scores where the score on one subtest of either MET or IELTS was more than two half-bands different from other subtests on that

same test, and where the same pattern of ability was not evident on the other test. For example, one participant's score on the Writing IELTS subtest was Band 2, while the scores on the Reading, Listening, and Speaking subtests were Band 4, Band 5, and Band 6.5, respectively; this participant's MET score for Writing, in contrast, was the highest among the MET subtests. This pattern indicates that the participant was not making an effort on the IELTS Writing subtest.

Falling outside the range was defined as participants whose overall IELTS scores were below Band 4.5 or above Band 8.5, as there were too few participants at these levels to provide meaningful data.

As a result of implementing these criteria, the following participants were removed from the study:

- 21 test takers whose overall MET scores were over 2SD higher than the band mean, and an additional 25 test takers whose overall scores were over 2SD below the mean.
- Five test takers with an anomalous performance on one of the subtests, a discrepancy by specific score. Of these, four people were removed from the dataset due to a discrepant Writing score and one person for a discrepant Reading score.
- Nine test takers with overall IELTS scores that fell outside the IELTS band score ranges: one test taker in Band 3, two test takers in Band 3.5, five in Band 4, and one in Band 9.

In addition to the removal of statistical outliers, six more participants were removed to improve the test order balance. The total number of participants removed was 66, and the final data set consisted of 1,000 participants.

Note that because the cleaning criteria applied to overall scores, the concordance tables for the subskills (see section 4.5) may still include IELTS section scores in bands outside the overall range.

For comparison of data distribution before and after the data cleaning, please see Figures 3.3.1 and 3.3.2.

Figure 3.3.1

Distribution of MET Scores by IELTS Bands, n = 1,066

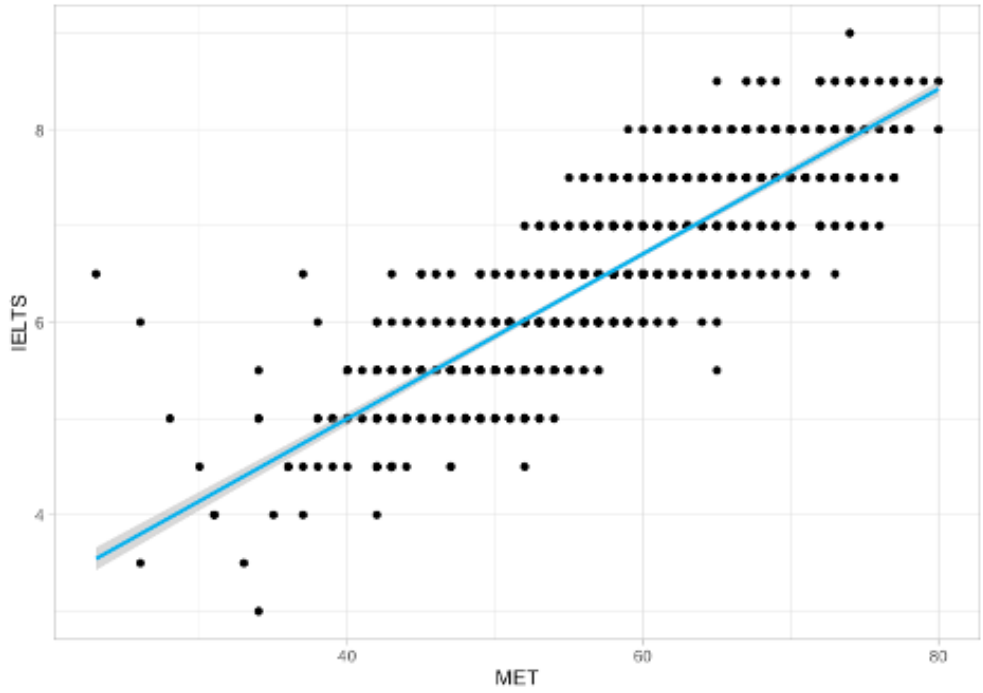


Figure 3.3.2

Distribution of MET Scores by IELTS Bands After Removing Outliers, n = 1,000

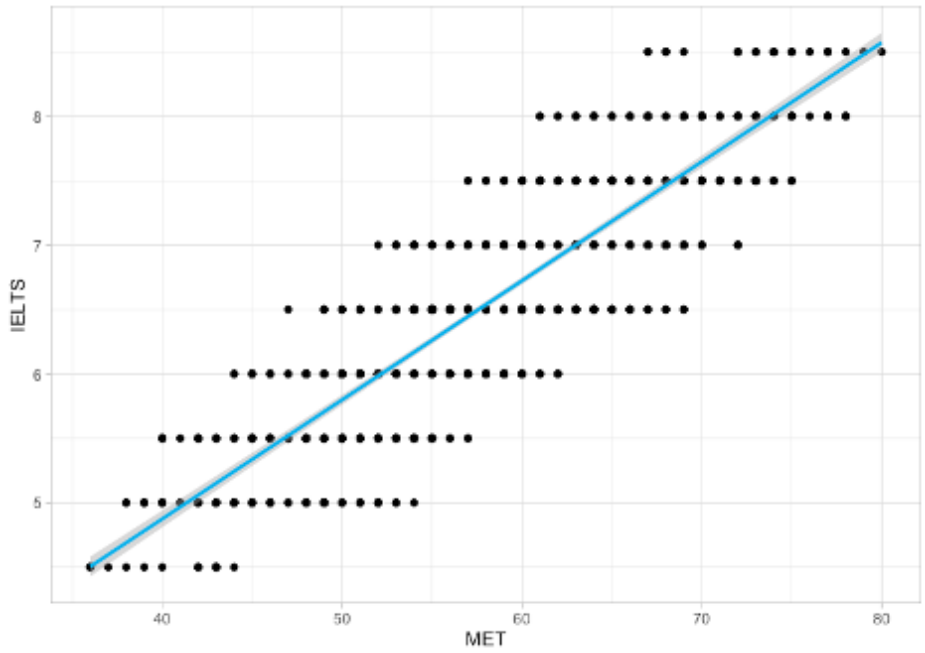


Table 3.3.2 reports the descriptive statistics of MET scores by IELTS band of the final sample used for linking after removing the outliers for a final sample of n = 1,000.

Table 3.3.2

Descriptive Statistics of MET Scores by IELTS Band After Removing Outliers, n = 1,000

IELTS band	n	mean	min	max	range	SD	SE
4.5	16	40.44	36	44	8	2.97	0.74
5	69	45.55	38	54	16	4.31	0.52
5.5	115	48.75	40	57	17	4.03	0.38
6	188	53.68	44	62	18	3.86	0.28
6.5	197	58.70	47	69	22	4.34	0.31
7	166	61.21	52	72	20	4.58	0.36
7.5	141	66.16	57	75	18	4.29	0.36
8	82	69.80	61	78	17	4.49	0.50
8.5	26	73.50	67	80	13	3.79	0.74

4.0 Results

4.1 Population

The population of interest for this study included international students interested in higher education studies in an English-speaking country, professionals interested in obtaining a work visa and securing employment in an English-speaking country, and individuals needing to show a level of English proficiency for domestic recognition purposes. Michigan Language Assessment's sampling procedures aimed at reaching a population that reflected a broad representation of nationalities. Test takers younger than 17 years of age were excluded from the study. Participants were required to take the two tests no more than three months apart.

4.1.1 Summary Demographics

Table 4.1.1.1

Summary of Participant Exam Taken Order

Exam Taken First	Count	Percent
IELTS	416	41.60
MET	584	58.40
Total	1000	100

Table 4.1.1.2

Summary of Participant Days Between Exams

Statistic	Value
Number	1000
Average	17.79
Standard Deviation	20.57
Minimum	0
First Quartile	3
Median	9
Third Quartile	28
Maximum	91

Table 4.1.1.3

Participant Gender Distribution

Gender	Count	Percent
Male	444	44.40
Female	554	55.40
Prefer to Self-Describe	1	0.10
Prefer Not to Answer	1	0.10
Total	1000	100

Table 4.1.1.4

Summary of Participant Age

Statistic	Value
Number	1000
Average	29.87
Standard Deviation	9.23
Minimum	17
First Quartile	23
Median	28
Third Quartile	36
Maximum	62

Figure 4.1.1.1

Participant Age Distribution

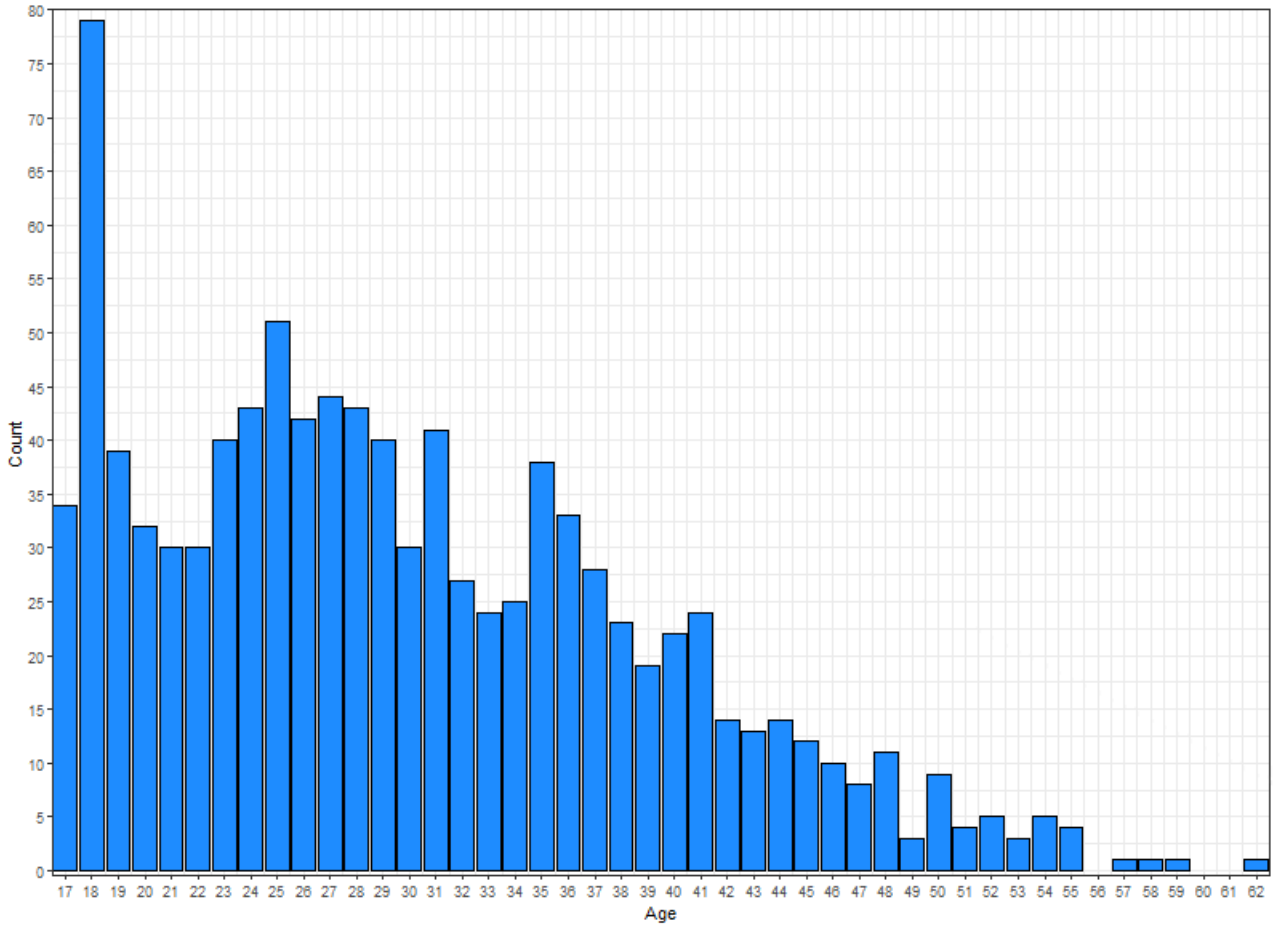


Table 4.1.1.5

Participant Education Level Distribution

Education Level	Count	Percent
Primary	5	0.50
Lower Secondary	2	0.20
Upper Secondary	185	18.50
Undergraduate	418	41.80
Postgraduate	390	39.00
Total	1000	100

Table 4.1.1.6

Participant First Language Distribution

First Language	Count	Percent
Afrikaans	2	0.20
Arabic	118	11.80
Bengali	66	6.60
Chinese (Cantonese/Mandarin)	39	3.90
Creole	1	0.10
English	81	8.10
Farsi/Persian	3	0.30
French	3	0.30
German	1	0.10
Greek	1	0.10
Gujarati	2	0.20
Hausa	2	0.20
Hindi	25	2.50
Ibo (Igbo)	7	0.70
Japanese	1	0.10
Korean	1	0.10
Malayalam	1	0.10
Nepali	87	8.70
Polish	1	0.10
Portuguese	60	6.00
Punjabi	3	0.30
Shona	1	0.10
Spanish	242	24.20
Swahili	1	0.10
Tagalog/Filipino	197	19.70
Thai	1	0.10
Turkish	6	0.60
Urdu	3	0.30
Vietnamese	2	0.20
Yoruba	21	2.10
Other	21	2.10
Total	1000	100

Of the 81 participants who listed English as their first language, most were from Nigeria and the Philippines, where English is an official language; however, because people from these countries may still be required to prove their English proficiency for immigration and educational purposes, it is legitimate to include them in the study. Participants selected language from a dropdown menu of the languages listed specifically in the chart, so that “Other” reflects participants’ selection of the “Other” category when their language did not appear on the dropdown list.

Table 4.1.1.7

Participant Country of Origin Distribution

Country of Origin	Count	Percent
Argentina	11	1.10
Bangladesh	66	6.60
Brazil	60	6.00
China	37	3.70
Colombia	137	13.70
Costa Rica	11	1.10
Egypt	111	11.10
India	31	3.10
Mexico	69	6.90
Nepal	87	8.70
Nigeria	79	7.90
Peru	12	1.20
Philippines	235	23.50
Other	54	5.40
Total	1000	100

Table 4.1.1.8

Participant Country of Testing Distribution

Country of Testing	Count	Percent
Argentina	11	1.10
Bangladesh	65	6.50
Brazil	60	6.00
Canada	35	3.50
Colombia	143	14.30
Costa Rica	11	1.10
Egypt	114	11.40
India	26	2.60
Mexico	72	7.20
Nepal	85	8.50
Nigeria	77	7.70
Peru	11	1.10
Philippines	241	24.10
United Kingdom	18	1.80
Other	31	3.10
Total	1000	100

.

4.1.2 Reason for Taking the Test

Participants in the study were asked why they were taking an English test. The options provided were education program admissions, nursing credentials, language course requirement, obtain a scholarship, obtain employment, improve employment, personal interest, volunteering, and other. For the purposes of Table 4.1.2.1, nursing credentials was included in “Employment” and education program admissions was counted as “Education,” while language course requirement, obtain a scholarship, volunteering, and personal interest were considered “Other.”

Table 4.1.2.1

Participant Reason for Taking Test Distribution

Reason	Count	Percent
Education	403	40.3
Employment	288	28.8
Other	309	30.9
Total	1000	100

4.2 Descriptive Statistics

Descriptive statistics for MET and IELTS samples are presented below in Table 4.2.1. The distribution of the scores is visualized in Figure 4.2.1, which shows that for both MET and IELTS, the participants in the study represented the full range of scores of interest for the study.

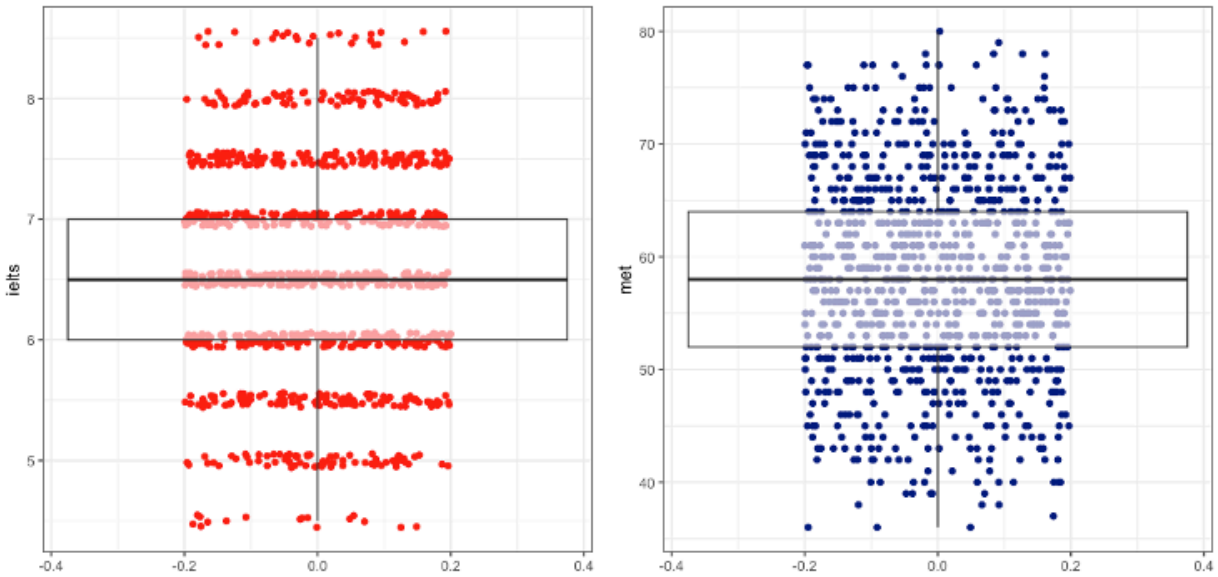
Table 4.2.1

Descriptive Statistics by Subtest

Test Statistic	IELTS			MET		
	Mean	Standard Deviation	Standard Error	Mean	Standard Deviation	Standard Error
Overall	6.55	0.92	0.03	58.18	8.70	0.28
Listening	6.72	1.34	0.04	60.10	9.33	0.30
Reading	6.55	1.25	0.04	59.60	10.54	0.33
Speaking	6.54	0.91	0.03	54.77	11.34	0.36
Writing	6.14	0.67	0.02	59.73	10.02	0.32

Figure 4.2.1

Distribution of Scores in the Sample (IELTS = left panel; MET = right panel)



4.3 Correlations Between Tests and Subtests

Pearson product-moment correlations were carried out to show the relationships between the subtest scores across the tests (n = 1,000).

Table 4.3.1

Subtest Correlations

		IELTS				
	Subtest	**Listening**	**Reading**	**Speaking**	**Writing**	**Total**
MET	Listening	**0.694**	0.717	0.646	0.610	
	Reading	0.692	**0.751**	0.683	0.638	
	Speaking	0.560	0.571	**0.654**	0.540	
	Writing	0.623	0.644	0.606	**0.642**	
	Overall					**0.872**

The tests exhibit moderate to strong correlations between the corresponding subtests, ranging from r = 0.642 for Writing to r = 0.751 for Reading. The correlation coefficient for the MET and IELTS total scores, r = 0.872, is in the “high” to “very high” range.

The robust nature of correlations between the observed test scores in the sample establishes the suitability of the sample for the concordance study. It should be noted that the correlation for IELTS Speaking is slightly higher for MET Reading than for MET Speaking; this may reflect in part the difference between the tests that IELTS Speaking requires the test taker to read, whereas MET Speaking does not require the reading skill at all.

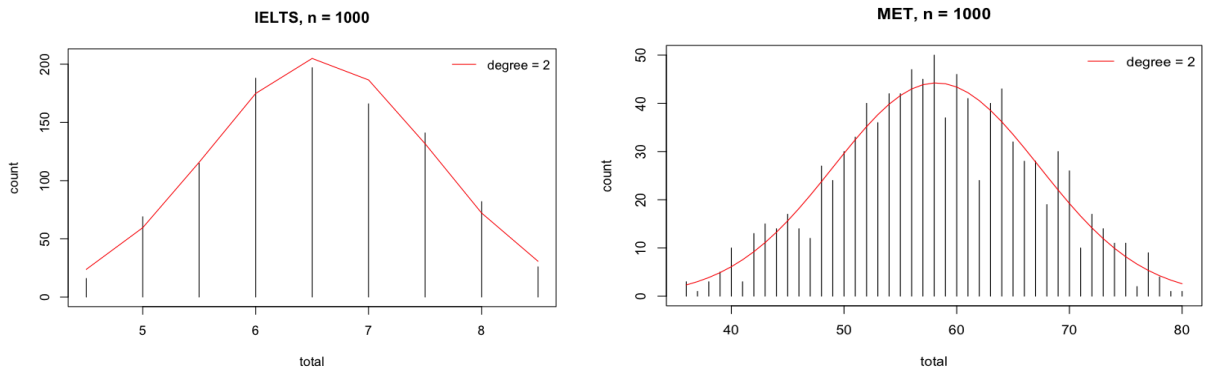
4.4 Equating Procedure

Test scores were subjected to polynomial log-linear presmoothing prior to linking. The method was selected as having been identified as statistically the most robust method in this context and was also used in previous IELTS linking studies (Elliot et al., 2021; Lim et al., 2013).

Several polynomial models were fitted to the data, and the Akaike Information Criterion (AIC) was used to select the best-fitting model, as well as the p-value of improvement compared to a simpler model. For both tests, 2-degree polynomial smoothing was selected as the best fitting model. For the IELTS sample, the 2-degree polynomial model had a better fit over the linear model (AIC = 73.2, $p < 0.001$), with the next best fit being a 4-degree polynomial model; however, the improvement was statistically significant only marginally (AIC = 70.54, $p = 0.015$). MET results followed the same pattern. The 2-degree polynomial was superior to both linear (AIC = 261.94, $p < 0.001$) and the 3-degree polynomial smoothing models (AIC = 263.88, $p = 0.810$, not a significant model improvement over a 2-degree polynomial model).

Figure 4.4.1

Best Fitting Presmoothing Model for Each Test

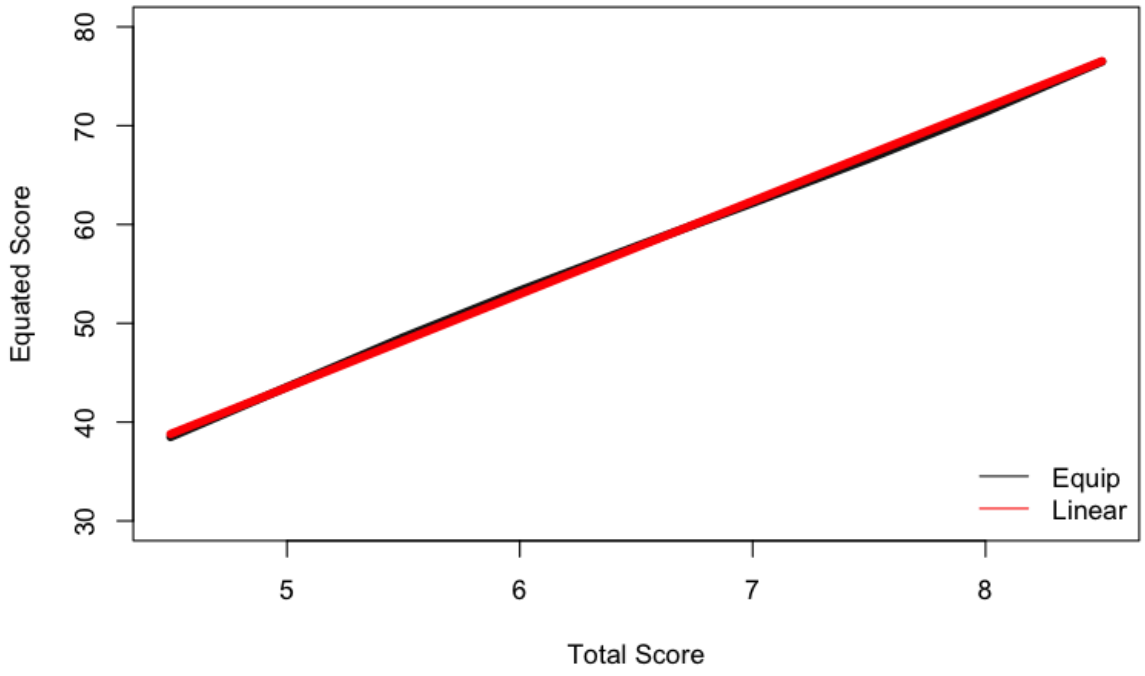


For linking we explored two different linking methods: linear and equipercentile, two methods commonly used for equating (Kolen & Brennan, 2014). We used the *equate* statistical package (Albano, 2016) running R version 4.3.2 (R Core Team, 2021).

As can be seen in Figure 4.4.2, the two methods produced very similar results. We report the results of the equipercentile method, as this method has been used in other concordancing studies with IELTS (see Clesham & Hughes, 2020; Cardwell et al., 2024) and, has the benefit of producing the same results regardless of which test is used as a reference point.

Figure 4.4.2

Equating Function for Total Scores, IELTS on x axis, MET on y axis.



4.5 Final Equivalence Tables

The tables below show the concordancing that was established for MET and IELTS scores. Each table shows the n (the number of study participants who had scores at that IELTS band) and the standard error (SE) at that band. As with the standard error of measurement (SEM) of the test as a whole, the SE at a given score level indicates how much variation in scores we could expect from a test taker at the given ability level if they took the test multiple times. In general, if the SE is lower than the SEM, that indicates that variations in the scores of the concordance will, practically speaking, not make a difference in how the scores can be interpreted.Therefore, if the SE of the concordancing is less than f the SEM of MET, the concordancing will be useful for most practical purposes. The results of the study showed that all the concordances are well within the SEM.

Overall concordances are presented first, followed by Listening, Reading, Speaking, and Writing. Note that the n column does not always total 1,000 for the specific skills, because although a few participants had IELTS band scores below 4.5, we are not including concordances for those scores. Note also that the top IELTS bands differ among the tables; as explained in Section 3.3, the sample was selected by IELTS overall score. Although no one in the sample had an IELTS overall score of 9, these scores are represented in the subskills for Listening, Reading, and Speaking; for Writing, the top score was IELTS 8.0.

4.5.1 Overall

Table 4.5.1

Overall at IELTS .5 Levels

IELTS band	MET range	n	SE (MET)
4.5	38 - 43	16	1.12
5	44 - 48	69	0.70
5.5	49 - 52	115	0.58
6	53 - 57	188	0.55
6.5	58 - 61	197	0.51
7	62 - 66	166	0.52
7.5	67 - 70	141	0.59
8	71 - 75	82	0.64
8.5	76 - 80	26	0.81

Generally, the equating produced reliable results across the IELTS scale; the Standard Error (SE) for each band is within a reasonable range, indicating that IELTS and MET are measuring similar constructs and are appropriate for linking. The higher SE for Band 4.5 can be attributed to a smaller number of test takers in the category; note that the SE value is still much smaller than the MET Standard Error of Measurement (SEM = 2.72), and therefore is still within the acceptable tolerance range for linking.

4.5.2 Listening

Table 4.5.2

Listening at IELTS .5 levels

IELTS band	MET range	n	SE (MET)
4.5	45 - 48	44	0.88
5	49 - 52	85	0.66
5.5	53 - 55	110	0.52
6	56 - 58	144	0.50
6.5	59 - 60	113	0.51
7	61 - 63	98	0.53
7.5	64 - 67	110	0.58
8	68 - 71	109	0.71
8.5	72 - 77	102	0.81
9	78 - 80	61	0.61

The equating for the Listening subtest produced reliable results across the IELTS scale; the Standard Error (SE) for each band is within a reasonable range, indicating that IELTS and MET

linking results are reliable and generalizable, and therefore the study outcomes demonstrate sufficient reliability of the linking of the listening subtests of MET and IELTS.

4.5.3 Reading

Table 4.5.3

Reading at IELTS .5 Levels

IELTS band	MET range	n	SE (MET)
4.5	42 - 46	50	1.02
5	47 - 50	81	0.72
5.5	51 - 54	126	0.63
6	55 - 58	159	0.59
6.5	59 - 62	159	0.60
7	63 - 66	113	0.66
7.5	67 - 71	105	0.76
8	72 - 75	55	0.88
8.5	76 - 78	88	0.76
9	79 - 80	44	0.39

The equating for the Reading subtest produced reliable results across the IELTS scale; the Standard Error (SE) for each band is within a reasonable range, indicating that IELTS and MET are measuring similar constructs in English reading proficiency and are appropriate for linking. Band 4.5 has a higher SE than the other bands (SE = 1.02); however, the SE value is still well within the SEM for the MET Reading section (SEM = 4.67), and therefore the study outcomes demonstrate sufficient reliability of the linking of the reading subtests of MET and IELTS.

4.5.4 Speaking

Table 4.5.4

Speaking at IELTS .5 Levels

IELTS band	MET range	n	SE (MET)
4.5	31 - 37	15	2.13
5	38 - 42	65	0.91
5.5	43 - 47	106	0.65
6	48 - 52	191	0.55
6.5	53 - 58	206	0.56
7	59 - 66	212	0.72
7.5	67 - 75	108	1.22
8	76 - 78	59	1.07
8.5	79	31	0.42
9	80	5	0.12

The equating for the Speaking subtest produced reliable results across most of the IELTS scale; the Standard Error (SE) for each band is within a reasonable range, indicating that IELTS and MET are measuring similar constructs in English reading proficiency and are appropriate for linking. Band 4.5 has a higher SE than the other bands (SE = 2.13); however, the SE value is still well within the SEM for the MET Speaking section (SEM = 5.16), and therefore the study outcomes demonstrate sufficient reliability of the linking of the reading subtests of MET and IELTS.

4.5.5 Writing

Table 4.5.5

Writing at IELTS .5 Levels

IELTS band	MET range	n	SE (MET)
4.5	38 - 44	19.00	1.95
5	45 - 50	64	0.76
5.5	51 - 56	192	0.52
6	57 - 63	289	0.51
6.5	64 - 73	273	0.69
7	74 - 78	106	0.96
7.5	79	50	0.37
8	80	7	0.10

The results of the equating for Writing subtests produced reliable results across most of the IELTS scale; the Standard Error (SE) for each band is within a reasonable range, indicating

that IELTS and MET are measuring similar constructs in English reading proficiency and are appropriate for linking. Band 4.5 has a higher SE than the other bands (SE = 1.95); however, the SE value is still well within the SEM for the MET Speaking section (SEM = 4.92), and therefore the study outcomes demonstrate sufficient reliability of the linking of the reading subtests of MET and IELTS.

4.6 Population Invariance

Evidence for population invariance was established through a population invariance study. Equating processes should be assessed with regard to the degree to which test scores and measurement properties remain the same across different population groups. A population invariance study provides evidence that item functioning and equating outcomes do not vary across groups defined by demographic or other background characteristics. The population invariance study for MET and IELTS linking evaluated possible differences in the population with respect to (1) gender, (2) first language background, (3) reason for taking the test, and (4) the order in which the two tests were taken.

The invariance studies were based on the general equating function calculated on the final cleaned sample, n = 1,000. The invariance studies included the entire sample, except for the gender invariance sample, in which one test taker selected “Prefer to self-describe” and one selected “Prefer not to answer” in the gender category; these participants were removed from the sample for the invariance study, for a total sample n = 998 (444 male, 554 female). Subpopulations for the invariance studies for the reason for taking the test and first language were grouped to achieve meaningful sizes of the subpopulation and include an “Other” category of individuals who did not fall into a meaningful group.

The invariance studies were based on Dorans and Holland’s methodology (2000), developed to measure the degree to which the two or more subpopulations’ equating outcomes are the same. The Root Mean Square Deviation (RMSD) compares the differences in equating functions as they pertain to separate populations to the function computed for the entire population, and can be interpreted as the effect size. The contribution of each population segment to the equating is weighted by its percentage in the overall sample. The RMSD metric is a percentage of the standard deviation of one of the tests’ measurement scales, which is dependent on the direction of the equating. Group sizes can affect the values of the RMSD, however, so RMSD values should be interpreted within the context of the specific test and its scale. Kolen and Brennan (2014) also note that the target values for RMSD will depend on the nature of the decisions being made for the particular test use. They propose that RMSD values of less than half of the test’s SD are appropriate. For MET in this context, we also compare RMSD expressed as MET scale score point values to the overall SEM of the test as another metric for interpreting RMSD values.

4.6.1 Gender Invariance Study

The table below lists the equating results for each IELTS score band and its corresponding MET score equivalent for Male, Female, and the entire population. The RMSD for the Male/Female invariance study is 0.039, the equivalent of less than 4% of a standard deviation for MET (SD = 8.698), or about 0.34 of a point on the MET scale, a negligible difference.

Figure 4.6.1

Equating Function by Gender

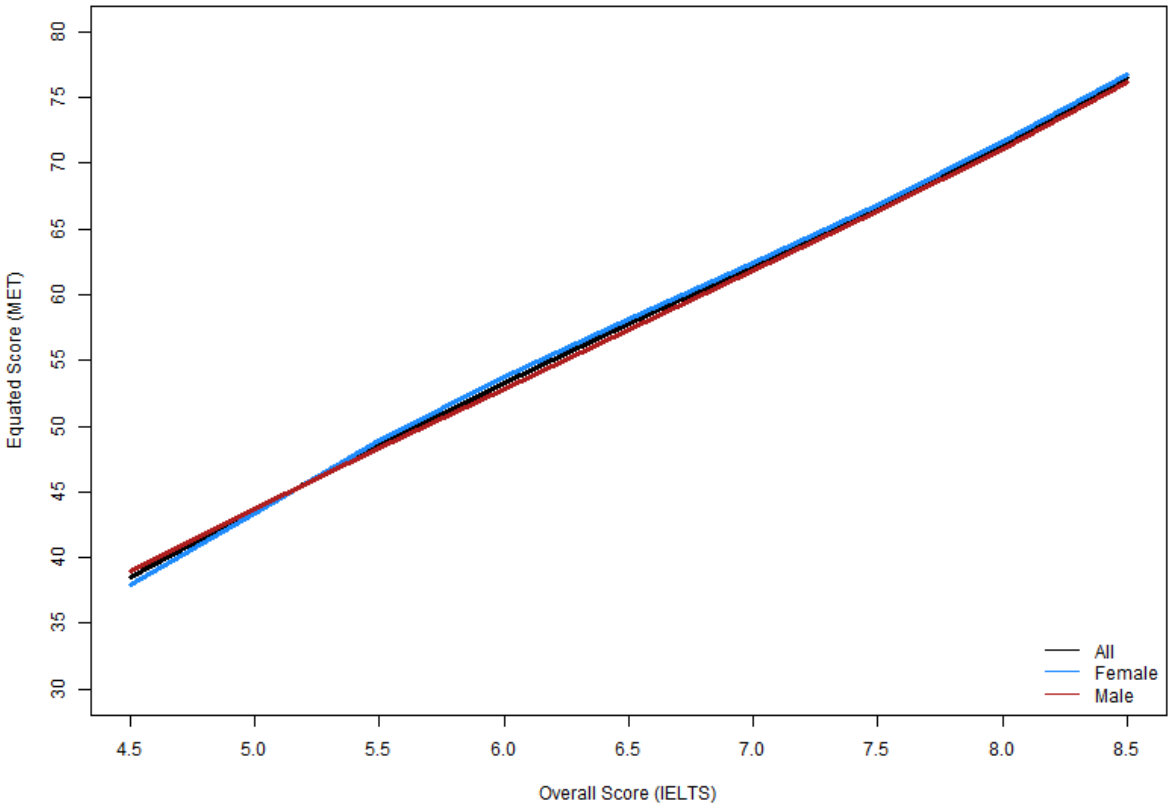


Table 4.6.1

Comparison of MET to IELTS Score Equating by Gender

Groups*	Male	Female	All
Count	444	554	998
Percent	44.40%	55.40%	99.80%
IELTS Band	**Male**	**Female**	**All**
4.5	39.03	37.89	38.49
5	43.71	43.35	43.55
5.5	48.29	48.86	48.58
6	52.83	53.77	53.31
6.5	57.34	58.18	57.79
7	61.83	62.44	62.17
7.5	66.37	66.87	66.66
8	71.06	71.64	71.40
8.5	76.20	76.79	76.48

*The subpopulations do not make up 100% because the two test takers who selected "prefer not to answer" and "prefer to self-describe" were excluded from this analysis.

4.6.2 Language Background Invariance Study

The language background invariance study compared major language groups represented by test takers in the sample (see counts and percentages in Table 4.6.2) to the total sample. Since MET is designed to test English in an international context, it is important to determine that the test is not biased toward speakers from particular families of languages. For the purposes of this study, we had languages from the Romance language family (Spanish, Portuguese, and French) and the Indo-Aryan family (many languages spoken in the Indian subcontinent, notably Hindi, Bengali, and Nepali), and in addition speakers of multiple dialects of Chinese, as well as Arabic and Tagalog.

There was more variability in the equating among the language groups than between the genders; however, the population invariance hypothesis was still upheld. The RMSD for this study is 0.171, the equivalent of 17.1% of a standard deviation for MET (SD = 8.698), or about 1.48 points on the MET scale. The overall SEM for MET is 2.72, close to double the value of the observed RMSD of 1.48. The fact that the RMSD is smaller than the test's SEM suggests a high level of population invariance and indicates that the differences between groups are generally negligible in practice and present no challenge to the generalizability of the equating outcome of the concordance study.

Figure 4.6.2
Language Background

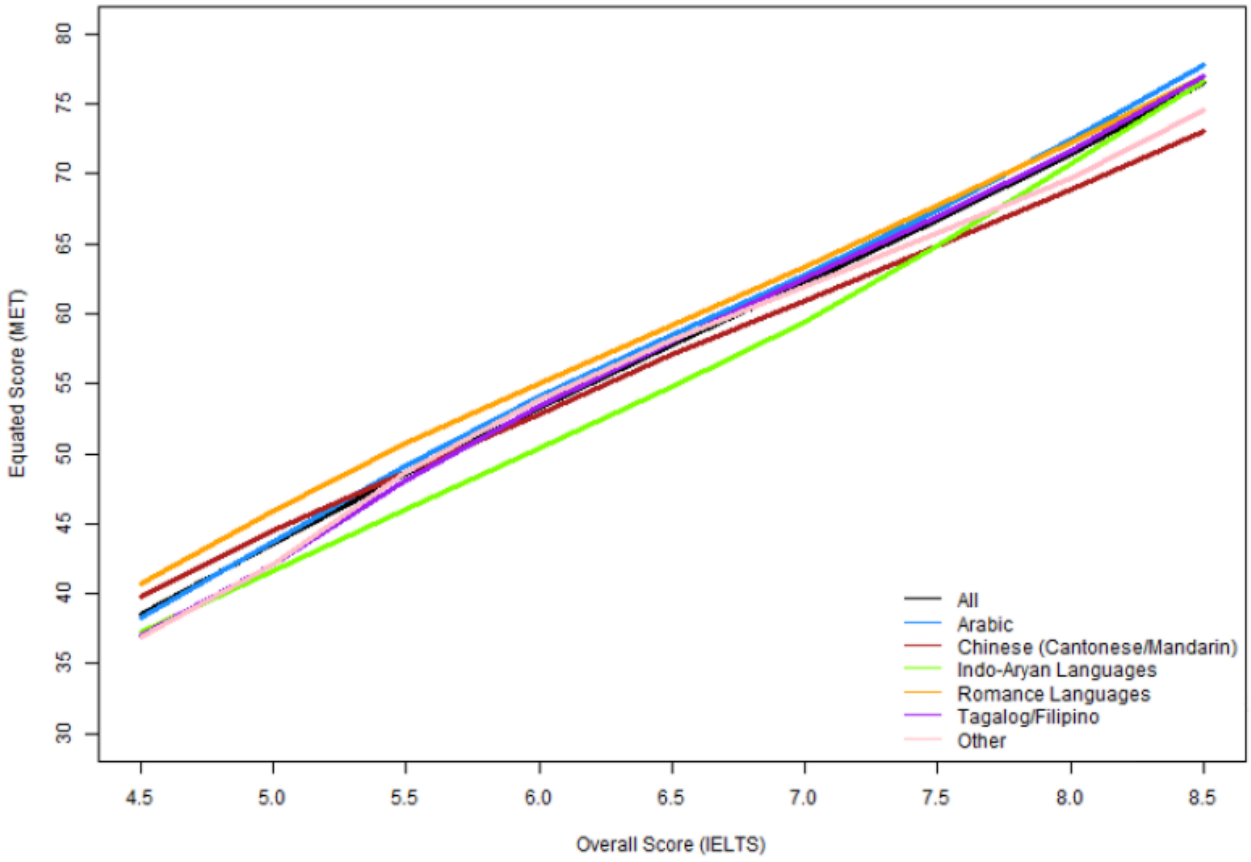


Table 4.6.2

Comparison of MET to IELTS Score Equating by Language Background

Groups	Arabic	Chinese (Cantonese/ Mandarin)	Indo-Aryan Languages	Romance Languages	Tagalog/ Filipino	Other	All
Count	118	39	187	305	197	154	1000
Percent	11.80%	3.90%	18.70%	30.50%	19.70%	15.40%	100.00%
IELTS Band	**Arabic**	**Chinese (Cantonese/ Mandarin)**	**Indo-Aryan Languages**	**Romance Languages**	**Tagalog/ Filipino**	**Other**	**All**
4.5	38.29	39.79	37.26	40.67	36.97	36.85	38.49
5	43.71	44.57	41.67	45.95	42.06	42.08	43.55
5.5	49.14	48.69	46.05	50.71	48.16	48.70	48.58
6	54.07	52.88	50.38	55.09	53.44	53.82	53.31
6.5	58.49	57.07	54.78	59.24	58.10	58.11	57.79
7	62.82	60.98	59.48	63.38	62.49	62.02	62.17
7.5	67.44	64.82	64.81	67.71	66.93	65.80	66.66
8	72.47	68.86	70.72	72.29	71.68	69.73	71.40
8.5	77.79	73.07	76.66	77.04	77.01	74.60	76.48

4.6.3 Testing Purpose Invariance Study

The testing purpose invariance study compared three groups to the total sample. The purpose for taking the test was reported by the study participants in the questionnaire they filled out prior to being accepted into the study. The categories in the testing purpose questionnaire were Education Program Admissions, Improve Employment, Obtain Employment, Language Course Requirement, Nursing Credentials, Scholarship, Volunteering, Personal interest, and Other (unspecified). The focus of the invariance study was the categories most relevant to high-stakes uses of the tests, which were those related to education and employment. Based on the collected responses, testing purpose was grouped into two categories important for high-stakes decision making: Education (Education Program Admissions, Language Program Requirements) and Employment (Improve/Obtain Employment, Nursing Credentials). Language Course Requirement, Scholarship, Volunteering, Personal Interest, and Other were grouped into Other, as these purposes are generally less high stakes and might potentially lead to different performances on the test.

Table 4.6.3 details the equating results for IELTS score bands and corresponding MET score population equivalent. The RMSD for this study is 0.081, the equivalent of 8.1% of a standard deviation for MET (SD = 8.698), or a difference of 0.71 of a point on the MET scale. The outcome of the invariance study falls within the permissible range of RMSD values for equating, is well within the test's standard error of measurement (SEM = 2.72), and, therefore, has no significant impact on the interpretation of the concordance results.

Figure 4.6.3

Testing Purpose

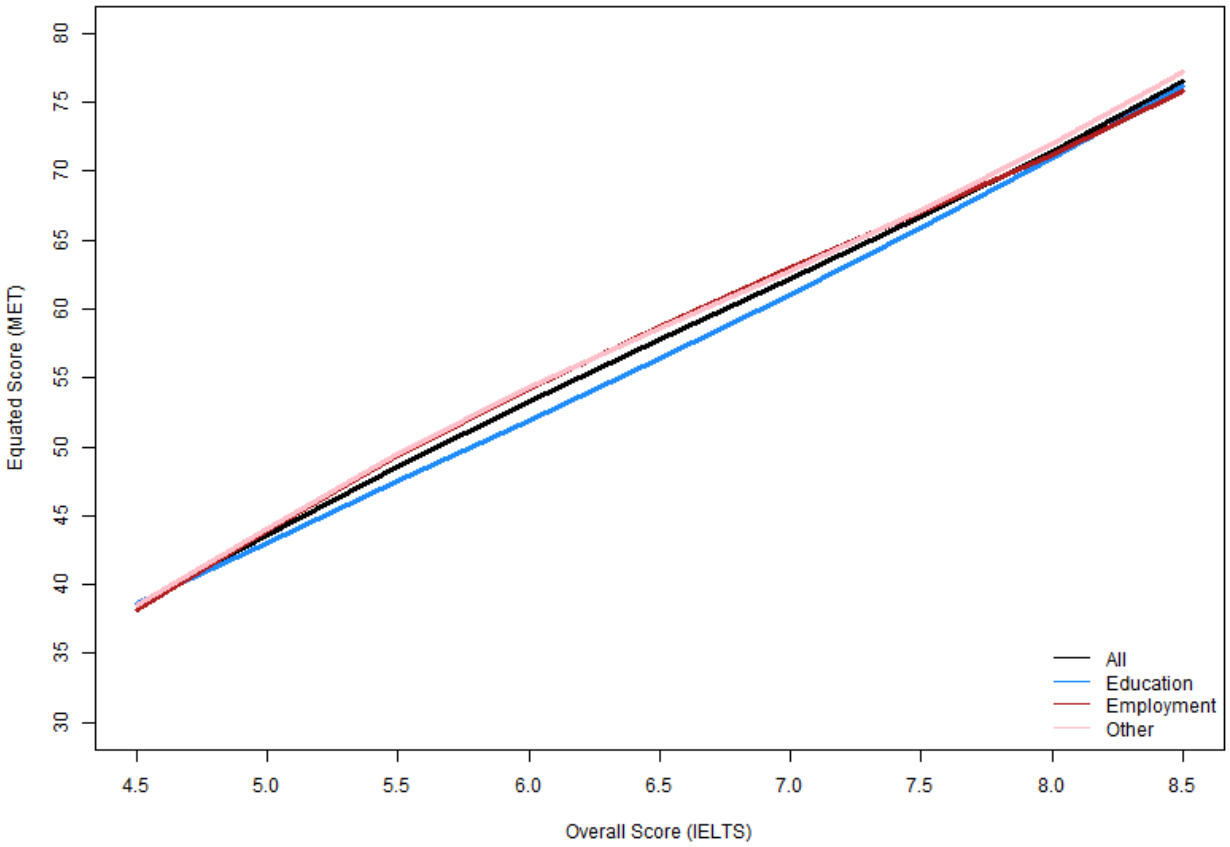


Table 4.6.3

Comparison of MET to IELTS Score Equating by Testing Purpose

IELTS Band	Education	Employment	Other	All
Count	403	288	309	1000
Percent	40.30%	28.80%	30.90%	100.00%
IELTS Band	**Education**	**Employment**	**Other**	**All**
4.5	38.58	38.20	38.51	38.49
5	43.06	43.90	44.09	43.55
5.5	47.54	49.41	49.52	48.58
6	51.97	54.28	54.28	53.31
6.5	56.43	58.74	58.59	57.79
7	61.05	62.97	62.75	62.17
7.5	65.93	67.07	67.13	66.66
8	71.01	71.20	71.97	71.40
8.5	76.14	75.87	77.18	76.48

4.6.4 Test Order Invariance Study

Michigan Language Assessment made every effort to counterbalance the order of test presentation to the test taker; however, due to logistical factors and test session availability, the order of presentation was skewed in favor of MET, with about 58% of study participants taking MET first, and 42% taking IELTS. The invariance study on test order explored the effect of the test order on the equating function.

The two equating functions—for the group who took IELTS first and for the group who took MET first—were compared against the total population function. The mean RMSD for this study is 0.042, the equivalent of 4.2% of a standard deviation for MET (SD = 8.698), or about 0.37 of a point on the MET scale, a negligible difference.

The presentation order of the test had no significant effect on the interpretation of the concordance results. Table 4.6.4 summarizes the results of the linking by test order, confirming that the results of the linking in the current sample are reliable and independent of the test order.

Figure 4.6.4

Test Order Presentation

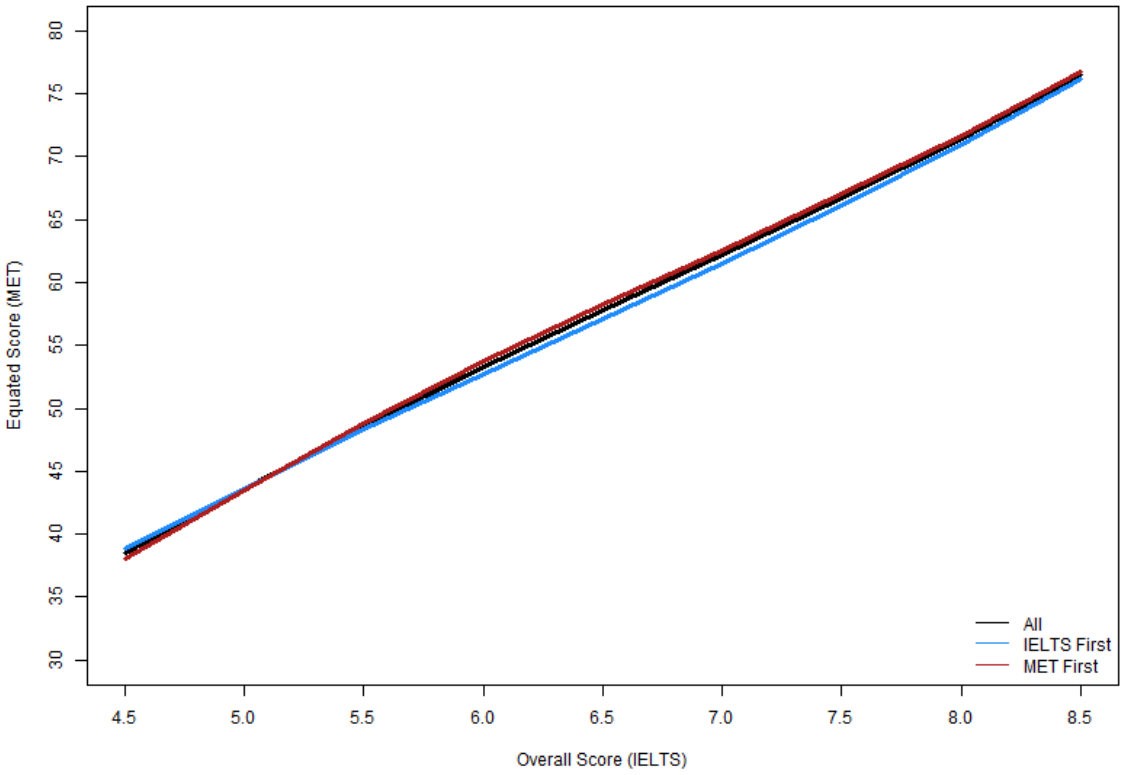


Table 4.6.4

Comparison of MET to IELTS Score Equating by Test Order Presentation

IELTS Band	IELTS First	MET First	All
Count	416	584	1000
Percent	41.60%	58.40%	100.00%
IELTS Band	**IELTS First**	**MET First**	**All**
4.5	38.92	38.11	38.49
5	43.65	43.47	43.55
5.5	48.28	48.83	48.58
6	52.77	53.72	53.31
6.5	57.16	58.22	57.79
7	61.56	62.56	62.17
7.5	66.14	66.99	66.66
8	70.98	71.69	71.40
8.5	76.17	76.75	76.48

4.6.6 Summary

The results of the population invariance studies support the linking between MET and IELTS and confirm that MET test score is not affected by gender, language background, the reasons for taking the test, or test order, and performs reliably in the populations represented in the study sample. These results confirm that the equating results can be replicated with a different proportion of the subpopulations in the sample, and the equating results are generalizable to a broader population.

5.0 Conclusions

Overall, the study results show that IELTS and MET are appropriately similar for the purposes for equating, and that the score equivalences are stable across subpopulations of interest. The sample population is sufficiently similar to the population of interest; each participant took the two tests within three months of each other; the score data was collected from official score reports; and correlations between overall and section scores are sufficiently strong. Although the number of participants at the very lowest end of the score scales is relatively small, the standard errors are generally within acceptable ranges for making high-stakes decisions.

5.1 Interpreting score comparisons

Score users—for example, institutions that use certain test scores for decisions about test takers—are advised that score comparisons across tests, while based on empirical research, are estimates only and should be treated with caution for the following reasons:

- Tests differ, sometimes significantly, in the ways information about English language ability is elicited and assessed. Score comparisons are only meaningful to the extent that the tests measure the same ability or skill.
- Tests often differ in the length of the reporting scales used (for example one test may report on a six-point scale and another on a 100-point scale). As a result, a one-to-one mapping of scores from one test to another is rarely possible.
- The choice of concordance study methodology may produce variations in results.
- The populations of test takers may differ (e.g., with respect to ages, nationalities, and language backgrounds of test takers) from the population used in the research that generated the score equivalences.
- The sample sizes used for comparing scores from different tests are generally small across all levels/ranges, especially at extreme ends of the scale.
- Score concordance results are generally more robust for proficiency levels with larger numbers of test takers.
- Large Standard Errors show that score equivalences are particularly imprecise at certain points on the ability scale.
- Because the score comparisons presented in the score comparison table are indicative only, score users are advised not to rely solely on published score equivalences in making their decisions. They should weigh evidence from additional sources where feasible.

5.2 Equivalence Summary

The table below is a quick-reference summary of overall score equivalence. For details on the overall score equivalence and skill-specific equivalence, see the tables in Section 4.5. As noted in Sections 3.3 and 4.5, the final study sample was limited to test takers who had an overall IELTS score between 4.5 and 8.5; however, some participants had scores of 9 in Listening, Reading, and/or Speaking, and no participants in the sample had an IELTS Writing score above 8.

Table 5.2.1

Overall Equivalence

IELTS band	MET range Overall	MET range Listening	MET range Reading	MET range Speaking	MET range Writing
4.5	38 - 43	45 - 48	42 - 46	31 - 37	38 - 44
5	44 - 48	49 - 52	47 - 50	38 - 42	45 - 50
5.5	49 - 52	53 - 55	51 - 54	43 - 47	51 - 56
6	53 - 57	56 - 58	55 - 58	48 - 52	57 - 63
6.5	58 - 61	59 - 60	59 - 62	53 - 58	64 - 73
7	62 - 66	61 - 63	63 - 66	59 - 66	74 - 78
7.5	67 - 70	64 - 67	67 - 71	67 - 75	79
8	71 - 75	68 - 71	72 - 75	76 - 78	80
8.5	76-80	72 - 77	76 - 78	79	-
9	-	78 - 80	79 - 80	80	-

References

Albano, A.D. (2016). “equate: An R Package for Observed-Score Linking and Equating.” *Journal of Statistical Software*, 74(8), 1–36.

Bachman, L.F. and Palmer, A.S. (1996) *Language Testing in Practice Designing and Developing Useful Language Tests*. Oxford University Press, Oxford.

Cardwell, R. L., Nydick, S. W., Lockwood, J. R., & von Davier, A. A. (2024). Practical considerations when building concordances between English tests. *Language Testing*, *41*(1), 192–202.

Clesham, R. & Hughes, S. R. (2020). 2020 Concordance Report: PTE Academic and IELTS Academic. London: Pearson. Retrieved from https://pearsonpte.com/wp-content/uploads/2020/12/2020-concordance-Report-for-research-pages.pdf

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

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume.* Council of Europe Publishing, Strasbourg.

Dorans, N. & Holland, P. (2000). Population invariance and the equatabilty of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*(4), 281–306.

Elliot, M., Blackhurst, A., O’Sullivan, B., Clark, T., Dunlea, J., & Saville, N. (2021). Aligning IELTS and PTE-Academic: A measurement study. In N. Saville, B. O’Sullivan & T. Clark (Eds.), IELTS Partnership Research Papers: Studies in Test Comparability Series, No. 2 (pp. 42–64). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

IELTS (2024a). IELTS and the CEFR. Retrieved from https://ielts.org/organisations/ielts-for-organisations/compare-ielts/ielts-and-the-cefr

IELTS (2024b). Test performance 2023–2024. Retrieved from https://ielts.org/researchers/our-research/test-statistics#Test_performance

Knoch, U., & Fan, J. (2024). Test score comparison tables: How well are they serving test users? *Language Testing*, *41*(3), 681–693.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Lim, G., Geranpayeh, A., Khalifa, H., & Buckendahl, C. (2013). Standard Setting to an International Reference Framework: Implications for Theory and Practice. *International Journal of Testing*, *13:* 32–49.

Michigan Language Assessment (2014). Linking the Common European Framework of Reference and the MET Writing Test, Michigan Language Assessment Technical Report, Michigan

Language Assessment. Retrieved from https://michiganassessment.org/wp-content/uploads/2014/12/MET-Writing-Linking-to-CEFR-20141106.pdf

Michigan Language Assessment (2024). MET 2023 Test Report. Retrieved from https://michiganassessment.org/wp-content/uploads/2024/11/MET-2023-Test-Report.pdf

Papageorgiou, S. (2010). Setting cut scores on the Common European Framework of Reference for the Michigan English Test, Michigan Language Assessment Technical Report, Michigan Language Assessment. Retrieved from https://michiganassessment.org/wp-content/uploads/2014/12/MET_StandardSetting.pdf

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org

Yu, G. (2021). IELTS Academic and PTE-Academic: Degrees of Similarity. In N. Saville, B. O'Sullivan & T. Clark (Eds.), *IELTS Partnership Research Papers: Studies in Test Comparability Series*, No. 2 (pp. 7–41). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

Appendix: Best Practices in Concordance Studies

Knoch & Fan (2024) outline several best practice principles that concordance studies should follow to be interpreted appropriately. Their criteria are listed in the table below, with comments on how the current study aligns with the criteria.

Good Practice Principles Summary

Criterion	Current Study
Tests measure closely related constructs	Yes; see section 2.2 for details.
Robust correlations between the two tests	Yes; overall correlation was 0.872, subskill correlations above 0.64. See section 4.3.
Test administration conditions are similar	Yes; both administered in secure, proctored environments. All MET tests are digital; a very small number of study participants took IELTS in its paper version. Note that although both tests have a remote-proctored online version, that administration condition was excluded from this study. See section 2.4.
Similar levels of reliability for overall test and subsections	Yes; the SEM values of the two tests are comparable after accounting for the discrepancies in the size of the scale between the two tests.See Section 2.7 for details.
Sampling of participants similar to population of interest	Yes; participants came from a broad variety of countries. Most participants reported taking a test for high-stakes purposes. See section 3.1.
Collect data on the participants' reasons for taking the tests	Yes; see section 4.1.2.
Comparable levels of familiarity with both tests	Yes; participants received information about both tests. Participants were recruited from institutions that, as part of their services, guide and assist their customers in preparing for standardized tests, including by running familiarization sessions. Michigan Language Assessment collaborated with the institutions to provide resources and support to help test takers become familiar with both tests. Web traffic statistics indicate that online resources for MET were accessed.
Ensure that data are based on official test score reports	Yes; all scores were obtained from official score reports for tests taken under secure conditions.
Adequate sample size	Yes; N = 1,000 (after removing outliers)

Criterion	Current Study
Counterbalanced order of testing	Adequately counterbalanced; 58% of the sample took MET first, then IELTS, and 42% took IELTS first, then MET. Population invariance established that test order did not affect equating (RMSD = 0.042).
Time between testing is sufficiently short	Yes; all participants took MET and IELTS within three months of each other, averaging less than one month. (See section 4.1.1.)
Population invariance	Yes; see section 4.6. All invariance studies (gender, language background, testing purpose, and test order) showed RMSD values within tolerance.
Study publicly available	Yes; the results of the study are publicly available.
Descriptive statistics and correlation coefficients	Yes; see section 4.2.
Describe statistical methods and procedures in sufficient detail	Yes; see sections 3.3, 4.4.
Report results for both overall test and subsections	Yes; see section 4.5.
Report sample size and standard error at each score level	Yes; see section 4.5.
Alert test users to exercise caution in interpreting and using concordance results	Yes; see section 5.1
Provide test user-focused guidelines and recommendations	Yes; see section 5.