The Duolingo English Test and East Africa: Preliminary linking results with IELTS & CEFR

Michel Bézy[*] **and Burr Settles**[†]

Abstract

We study the validity of the Duolingo English Test for high-stakes university admissions, using data from Carnegie Mellon University's satellite campus in Rwanda. We find that Duolingo scores submitted during the 2015 application process are significantly correlated with IELTSTM scores ($r = 0.834$). We then link Duolingo test scores to the IELTS exam, with results that are consistent with previous research linking Duolingo scores to the TOEFL® iBT exam. Finally, we infer CEFR (Common European Framework of Reference) levels from this linking, and find that they are consistent with the design of the Duolingo English Test scale.

Keywords

Duolingo English Test, CEFR, English proficiency, IELTS, language testing

---

This report refers to an older version of the Duolingo English Test (DET). More recent research and information on the current test can be found here.

1 Introduction

In 2011, the government of Rwanda strategically targeted Carnegie Mellon University to establish and operate an academic program in Kigali, Rwanda. The program offers Master of Science degrees in Information Technology and in Electrical and Computer Engineering. These degrees are taught to the same rigorous standards as the main CMU campus in Pittsburgh, by globally renowned CMU faculty residing on-site in Kigali. As a result, the admissions requirements for these programs are similar to those in Pittsburgh, and currently require scores from either the IELTS[TM] or TOEFL® as part of the application process[1].

Experience has shown that mastery of English is a key success factor for students applying to these programs. Therefore, the CMU Rwanda admissions committee needed an efficient and affordable way to test students' English ability early in the admissions process, before requiring them to take the IELTS or TOEFL exams. Mainly, this is because because such tests are time-consuming and expensive to use in the admissions cycle: they must be scheduled in advance, there are a limited number of slots available at physical test centers, results might take weeks, and they can cost up to 34% of the annual per capita income in Rwanda[2].

In 2014, Duolingo — creator of the award-winning online language-learning app — launched the Duolingo Test Center (https://testcenter.duolingo.com) as a new online language certification platform. Duolingo Test Center offers secure and accessible computer-adaptive language assessments, on-demand, via web and mobile devices. The Duolingo English Test (DET) costs US$50, is proctored remotely using the device's built-in camera and microphone, and results are available in less than 48 hours. Furthermore, previous research has found that Duolingo English Test scores are substantially correlated with the TOEFL iBT (Internet-based test), providing early validity evidence for its use in university admissions decisions [1].

CMU Rwanda elected to use the Duolingo English Test as part of the admissions process for its 2015 applicants. The goal of this case study is to better understand how the Duolingo English Test can be used for university admissions decisions. We do this by analyzing validity evidence using data collected during the application process.

2 Data and method

2.1 Data collection

Students from the East African Community region who apply to CMU Rwanda degree programs must first submit an online application form. They are then required to submit a Duolingo English Test and perform two interviews. The most competitive students are then selected based on their DET scores and interviews — these students are considered to be in the top 5-10% of their class. If selected, students are then required to take the GRE and either the IELTS or TOEFL exams. A final admissions decision is made based on all of the information collected during the application process.

Data were obtained from students applying to CMU MS programs, who have already completed an undergraduate degree in a technology-related field. Students were all from East Africa: specifically Rwanda (89%), Uganda (8%), and Kenya (3%). The gender distribution was 71% male and 29% female, with an age range of 23–35 years (median 27).

---

[*]Associate Director, Carnegie Mellon University, Kigali, Rwanda
[†]Staff Research Scientist & Software Engineer, Duolingo, Pittsburgh PA, USA

This report was updated August 10, 2016 to incorporate changes to Duolingo Test Center since the manuscript was originally published. No results or conclusions are affected. First, the official name of the test was changed to "Duolingo Englist Test (DET)." Second, at the time our data were collected, DET scores were reported on a scale from 0–10. Duolingo switched to a 100-point scale in September 2015 (simply multiplying previous scores by 10). This report was updated to reflect these changes.

As a result of the admissions process, CMU collected data that allows Duolingo scores to be compared to other English assessments, for students who passed the second selection step. Among the 39 applicants who met the criteria, 36 submitted IELTS scores, 1 submitted TOEFL scores, and 2 submitted both. Given the low number of TOEFL submissions — presumably due to its higher cost in the area — we restrict our analysis to the Duolingo English Test and IELTS, which were taken about one month apart.

2.2 Duolingo, IELTS, and CEFR details

The Duolingo English Test (DET) is a computer-adaptive test of general English language ability. A major advantage of computer-adaptive tests is that they can provide uniformly precise scores for the full range of test-takers, by administering more of the difficult items to high-ability students, and more of the easy items to low-ability students. Such tests can typically be shortened by 50% or more (compared to a fixed, non-adaptive version) and still maintain this higher level of precision [2]. The DET lasts anywhere from 10–25 minutes (median 17 minutes). It uses interactive test item formats that take advantage of modern speech technologies and the camera/microphone now available on most computers and mobile devices. These sensors both enable the interactive exercises and record a video of the test-taker, which is reviewed by a human proctor to ensure test integrity. Scores are reported on a scale of 0–100. The turnaround time for proctoring and results is less than 48 hours (median 18.5 hours).

The International English Language Testing System (IELTS) is a joint effort of the British Council, IDP, and Cambridge English. It is commonly used to assess English language proficiency for people who want to study or work where English is the primary language of communication. The test is composed of four *modules*, each targeting a specific skill: reading, listening, speaking, and writing. The speaking module is a face-to-face interview with a certified examiner, and the writing module includes two short essay tasks that are scored by human raters. The reading and listening modules are automatically scored, using a variety of typically paper-and-pencil formats: multiple choice, short-answer, diagram-labeling, etc. Each module is scored from 1 to 9 (in increments of 0.5), with the rounded average producing an *overall band score*. The exam takes about 2.75 hours, and results are typically available after two weeks.

The Common European Framework of Reference (CEFR) is a guideline used to describe learner achievements for languages across Europe (and increasingly in Africa and other continents). It was proposed by the Council of Europe to provide a generic framework for learning, teaching, and assessing all languages [3]. The framework divides learners into three broad categories that are subdivided into six ability levels: A1 and A2 (basic users), B1 and B2 (independent users), C1 and C2 (proficient users). The demands of most English-language universities require a B2 or higher to be academically successful. The Duolingo scale was originally designed to correspond to the CEFR by mapping from $0 \approx A1$ to $100 \approx C2$ in roughly 20-point increments. The IELTS scale, while not based on CEFR levels, has been linked to them [4]. We discuss these relationships more in Section 3.3.

2.3 Data analysis

We provide descriptive statistics of all exam scores, and relate DET scores with IELTS scores using Pearson and Spearman correlation. We also use the circle-arc linking method [5] to produce a preliminary partial concordance that maps DET scores to IELTS overall band scores and, by extension, putative CEFR levels (since the Duolingo scale was based on CEFR levels, this linking provides some validation of Duolingo's scale design). We conduct all statistical analyses using the R platform [6], using the equate package [7] for circle-arc linking. All data were anonymized before any analysis took place.

3 Results

3.1 Descriptive statistics of test scores

Table 1 presents the mean (μ), standard deviation (σ), and range for both Duolingo and IELTS scores. For comparison, we also provide the mean IELTS scores from the 2013 test-taker population [8], the most recent year with data available (data specific to East Africa are not publically available).

Table 1. Summary of sample test scores ($N = 38$).

Test	μ	σ	Range	μ 2013
Duolingo	55.0	24.5	4–100	—
IELTS: Overall	5.8	0.8	4.5–7.5	5.9
Reading	5.7	1.0	4–8.5	5.9
Listening	5.5	1.2	3.5–8	6.0
Speaking	6.3	0.8	5–8.5	5.9
Writing	5.8	0.6	4.5–7	5.5

We see that the mean Duolingo and IELTS overall band scores are similar (both centered around the midpoint of the range), although the Duolingo scores have more spread. Compared to the 2013 population, sample IELTS overall band scores are nearly identical. Reading and listening scores appear to be slightly lower than the population, while speaking and writing scores appear slightly higher. Differences in gender or age show no significant effect on test scores.

3.2 Correlation between Duolingo and IELTS

A common type of structural validity evidence used in assessments — known as *criterion-related validity* — answers the question, "how well does this assessment predict current or future performance on some other trusted criterion?" If we consider IELTS to be a trusted criterion for admission to English-language universities, then a strong correlation between Duolingo and IELTS scores provides external structure evidence for using the Duolingo test in the same capacity. Table 2 presents Pearson (linear) and Spearman (rank) correlation coefficients, which measure the relationships between Duolingo scores and IELTS scores (overall band scores as well as skill module scores).

All correlations are statistically significant ($p \ll 0.001$)[3]. Pearson and Spearman correlations appear consistent, which is important because linearly-related scores may not necessarily produce similar rankings (and vice versa); considering one but not the other can be misleading [9]. The relationship between tests appears to be strongest for overall English ability and receptive language skills (reading and listening). Figure 1 shows
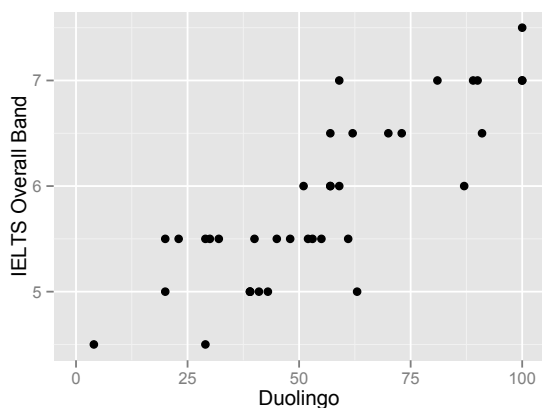
Figure 1. Scatterplot of Duolingo and IELTS scores.

Table 2. Correlations between Duolingo and IELTS scores.

IELTS Score	Pearson r	Spearman ρ
Overall	0.834	0.809
Reading	0.815	0.800
Listening	0.760	0.738
Speaking	0.675	0.688
Writing	0.584	0.610

a scatterplot of Duolingo and IELTS overall band scores, where the relationship is clearly visible.

3.3 Preliminary linking results

For two correlated assessments, it is useful to have a linking that can be used to map scores from one test to the other for concordance purposes. This section presents preliminary linking results that relate Duolingo scores to IELTS overall band scores based on our data[4]. Since previous research has linked IELTS scores to CEFR levels [4], we leverage this to infer a putative partial mapping from Duolingo $\rightarrow$ IELTS $\rightarrow$ CEFR levels, by joining both concordance tables.

While the correlation coefficients in Table 2 are slightly lower than the 0.866 sometimes recommended for concordance (based on the argument that one assessment should explain at least 75% of the variance in the other's score) [10], they are still substantially higher than the correlations used for many language assessment linking studies, which are typically in the 0.67–0.76 range [1, 11, 12].

Table 3 provides a partial concordance table using the circle-arc linking method[5]. For example, this table predicts that if a student obtains a Duolingo English Test with a score of 75, she should receive an IELTS overall band score of 6.5, and be able to communicate at CEFR level B2. Since our sample does not cover the full range of IELTS scores (and fewer than 5% of our subjects scored either a 4.5 or 7.5 on the IELTS), we only link Duolingo scores to IELTS scores ranging from 5 to 7.

Two interesting observations. First, a Duolingo score of 75 corresponds to 6.5 on the IELTS, which is a common cutoff for admission to selective U.S. universities. Previous results showed that a Duolingo score of 75 corresponds to 102 on the TOEFL iBT [1], which is very close to the common selective university cutoff of 100 for that test. This confluence provides additional evidence that the Duolingo English Test is at least as trustworthy as these other tests for making admissions decisions.

Table 3. Preliminary partial concordance between Duolingo scores, IELTS overall band scores, and CEFR levels.

Duolingo	IELTS	CEFR
25	5	B1
43	5.5	B2
60	6	B2
75	6.5	B2
89	7	C1

Second, as mentioned in Section 2.2, the Duolingo scale was designed to correspond to the CEFR such that $40 \approx$ B1, $60 \approx$ B2, and $80 \approx$ C1. This specification is consistent with CEFR levels that can be inferred from the linking in Table 3. This provides structural validity evidence for at least this portion of the Duolingo's scale design.

4 Conclusion

From these results we can draw several inferences:

1. Duolingo English Test scores are significantly correlated with IELTS scores ($r = 0.834$), providing additional criterion-related validity evidence for the Duolingo English Test in assessing English ability.

2. In a preliminary linking from Duolingo to IELTS, a 75 Duolingo score corresponds to a common admissions cutoff at selective U.S. universities. This is consistent with previous results linking to TOEFL iBT scores.

3. CEFR levels that can be inferred from Duolingo $\rightarrow$ IELTS $\rightarrow$ CEFR linkings are consistent with what the Duolingo scale specification predicts, providing additional validity evidence for its design.

All of these results suggest that the Duolingo English Test is a valid and effective instrument for evaluating English language proficiency in high-stakes university admissions. In addition, Duolingo Test Center is available on-demand, provides scores in less than 48 hours (18.5 on average), and the test is about 10% of the cost of IELTS.

Limitations

While our data were collected in an actual high-stakes admissions setting, and the two tests were administered very close together in time, there are a few limitations to the generalizability of these results. First, our sample size is 38 subjects, whereas 100 or more are preferred for reliability and concordance purposes [13]. Also, a few students passed the second admissions step based on their interviews alone and no Duolingo English Test (see Section 2.1), due to technical difficulties using the Duolingo Test Center website over unreliable Internet connections. These students are omitted from our analysis. While there is no reason to believe that this introduces a systematic bias into our data, it remains a practical concern for many use cases of the Duolingo test. (Note that Duolingo is continuously optimizing the service for low-bandwidth networks). Second, the population of subjects is relatively restricted: they are all from East Africa, applying to a specialized program at the same university, who had passed an initial selection process which was partially based on Duolingo scores. As such, results should be viewed as strong additional evidence in support of Duolingo English Test score validity, but not necessarily conclusive for all applications (pending a larger study with more statistical power).

Notes

1. The IELTS (International English Language Testing System) and TOEFL (Test of English as a Foreign Language) are English language assessments commonly required for admission of non-native English speakers to English-language colleges and universities.

2. In Rwanda, IELTS and TOEFL administrations cost US$200 and US$250, respectively. The 2014 per capita income was US$741 [14].

3. These correlations are also substantially higher than previous results relating Duolingo to TOEFL iBT scores [1]. Note, however, that those results were based on a larger and more varied population than the present study, and time between test administrations was less controlled.

4. These results are "preliminary" in the sense that our data set is relatively small ($N = 38$) and does not cover the range of all possible test scores.

5. Circle-arc is the preferred linking method for small sample sizes (< 100 common subjects) [5]. We also performed linear and equipercentile linking, but circle-arc produced the lowest bootstrap error estimates on our data.

References

[1] F. Ye. Validity, reliability, and concordance of the Duolingo English Test. Technical report, University of Pittsburgh, May 2014.

[2] D.J. Weiss and G.G. Kingsbury. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21:361–375, 1984.

[3] Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[4] The British Council. Relating IELTS scores to the Council of Europe's Common European Framework of Reference (CEFR). http://bit.ly/1D2gVIJ, retrieved July 2015.

[5] S.A. Livingston and S. Kim. The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46:330–343, 2009.

[6] R Core Team. R: A language and environment for statistical computing. http://www.r-project.org/.

[7] A. Albano. equate: Observed-score linking and equating. https://cran.r-project.org/web/packages/equate/.

[8] IELTS Researchers. Test taker performance 2013. http://bit.ly/1vvMOWv, retrieved July 2015.

[9] J. Hauke and T. Kossowski. Comparison of values of Pearson and Spearman correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 30(2), 2011.

[10] N.J. Dorans. Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4):227–246, 2004.

[11] Educational Testing Service. Linking TOEFL iBT scores to IELTS scores—a research report. 2010.

[12] Pearson. Preliminary estimates of concordance between Pearson Test of English Academic and other measures of English language competencies. 2009.

[13] G.A. Johansen and G.P Brooks. Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement*, 70(3):394—400, 2010.

[14] International Monetary Fund. World economic outlook database. http://bit.ly/1zh71Mk, retrieved July 2015.