Validity, reliability, and concordance of the Duolingo English Test

May 2014

Feifei Ye, PhD

Assistant Professor
University of Pittsburgh
School of Education

feifeiye@pitt.edu

Executive Summary

Duolingo has developed a computer adaptive test of English competency for non-native English learners. This research study of the validity and reliability of the Duolingo English test was independently conducted from February-April of 2014. The study lasted for approximately eight weeks. Participants were recruited from Duolingo users who studied English, international students in several United States universities, and people who took Test of English as Foreign Language (TOEFL) at several TOEFL centers in China. Participants were at least 18 years of age and had taken the TOEFL within the last 18 months.

Participants filled out a survey in the beginning of the study to provide information of their demographics such as age, gender, native language, and education background, experience in learning English, and most importantly, the date and location of taking the TOEFL iBT as well al their TOEFL scores. The participants were directed to complete the Duolingo English test at the end of the survey. Two weeks later, the participants were instructed via email to fill out a second survey and the Duolingo English test for the second time.

Main Results

- Participants' scores on the Duolingo English test correlated substantially with their TOEFL total scores, and moderately with TOEFL section scores, with higher correlations for the TOEFL Speaking and Writing sections. This provides criterion-related evidence of validity of the Duolingo test scores.

- Participants' scores on the first Duolingo English test correlated highly with their scores on a second test two weeks later, indicating that Duolingo test scores are reliable with a good test-retest reliability coefficient.

- Duolingo English test scores were linked with TOEFL total scores to find the comparable scores from the two tests that have similar percentile rank. Duolingo English test scores are on a scale of 0-10 and TOEFL scores are on a scale of 0-120. For international students to apply for studying in US universities, the minimum cut-off score of TOEFL iBT is 80 and a more selective cut-off score is 100, corresponding to scores 5.0 and 7.2 respectively on the Duolingo English test.

1. Introduction

Duolingo is a free language-learning website which became publicly available in 2012. Duolingo has recently developed a computer adaptive test of English competency for non-native English learners. The goal of the current research study was to independently evaluate some of the evidence for the validity and reliability of Duolingo English test scores, and to link Duolingo English test scores to TOEFL iBT scores. This study was designed and carried out to answer the following three research questions:

1) What is the external structure evidence for validity of scores on the Duolingo English test?

2) What is the test-retest reliability of scores on the Duolingo English test?

3) What scores on the Duolingo English test are comparable to TOEFL iBT total scores?

2. Method

2.1 Data Collection

To recruit subjects for this study, a banner was posted on the Duolingo website for Duolingo users learning English, and advertisements in Spanish and Portuguese were also placed on Google and Facebook. In addition, emails were sent to several TOEFL centers around the world, and the office of international studies at several universities within the United States. Eligibility for participation in this study was that the subject must be at least 18 years old and have taken the TOEFL after October 1, 2012 so that there was at most an 18-month interval before subjects took the Duolingo English test for the first time.

The advertisement and emails included a link to a survey developed by the researcher on Qualtrics. At the end of the survey, subjects were directed to a website developed by Duolingo for administering the first Duolingo English test. In the survey and the test page, the subjects were asked to fill out a valid email address which was used to link their survey responses and test scores. Two weeks after subjects took the first Duolingo test, the researcher emailed these subjects to invite them to fill out another survey and again, at the end of the survey, subjects were directed to the Duolingo test website to complete the test for the second time. The data collection lasted for about two months, and after all data were collected, Duolingo provided test details and scores to the researcher for each email address.

The first survey included questions regarding demographics and brief English learning experience. Most importantly, subjects were asked to report the date and the city where they took the TOEFL, to fill in the TOEFL section scores, and to provide an electronic copy of their official TOEFL score report. The second survey included questions regarding subjects' impressions of the Duolingo English test and will not be discussed in this study.

The Duolingo English test aims to assess proficiency in daily English of non-native English language learners. It was developed as a computer-adaptive test with four kinds of test items to assess reading, writing, listening, and speaking abilities, as presented in Appendix 1. The items selected for each subject differ with each administration, and depend on the subject's English proficiency (e.g., subjects with better English skills receive more difficult test items). Thus, the number of items varies as well for different subjects. The average test takes about 16 minutes, with a maximum length of 20 minutes.

2.2 Sample Description

An initial pool of 258 subjects was identified as they reported nonzero TOEFL scores and completed at least one Duolingo English test. A further examination of the TOEFL scores showed that 21 subjects reported scores on the TOEFL iBT taken more than 18 months ago and 5 subjects reported scores on the TOEFL ITP (a different test). An examination of Duolingo English test scores identified another 18 subjects with invalid responses. Invalid responses included selecting only gibberish words for vocabulary items, typing the same answer for all listening items, or having incorrect responses to more than 7 consecutive test items. These participants were considered as not motivated to properly complete the test. All these 44 subjects were excluded. The final sample consists of 214 subjects, and among them 165 subjects uploaded an official copy or screenshot of their TOEFL score report.

Table 1 presents demographic information of the final sample. The three most frequent native languages of the final sample are Chinese (46.3%), Spanish (33.6%), and Portuguese (6.1%). The three most frequent device languages of Duolingo English users are Spanish (33%), Chinese (22%), and Portuguese (16%). While the study sample has the same most frequent native languages as the Duolingo users, this study oversampled subjects speaking Chinese. The countries with more than 10 subjects are shown in the table and the country with the two largest frequencies are China (46.3%) and India (7.9). The top two countries where Duolingo English users come from are China (26%) and Brazil (18%). Again, compared to Duolingo users, the study oversampled Chinese. The majority of subjects in this sample (71.5%) were college students, and had education level higher than some college (94%). About 41% have studied in the United States.

Table 1. Demographic and background of subjects

		n	%
Gender	Male	108	50.5
	Female	106	49.5
Country originally from	China	99	46.3
	India	17	7.9
	Columbia	15	7.0
	Mexico	15	7.0
	Guatemala	14	6.5
	Chile	13	6.1
	Brasil	12	5.6
	Other	29	13.6
Education level	MA or higher	26	12.1
	Some graduate school	39	18.2
	Bachelor's degree	67	31.3
	Some college	69	32.2
	High school diploma or GED	13	6.1
Native language	Chinese	99	46.3
	Spanish	72	33.6
	Portuguese	13	6.1
	Other	30	14.0
Employment status	Full-time student	153	71.5
	Unemployed	26	12.1
	Full-time employed	17	7.9
	Part-time employed	13	6.1
	Other	5	2.3
Have studied in the US	School	145	67.8
	Personal interest	36	16.8
	Business/work	19	8.9
	Travel	4	1.9
	Other	10	4.7
	Yes	88	41.1
	No	126	58.9
		Mean	Standard Deviation
Years learning English		11.26	5.39
Age		23.92	3.35

2.3 Data Analysis

For the first research question, the Duolingo test scores from 214 subjects were correlated with their TOEFL iBT total and section scores using Pearson correlation. For the second research question, 107 subjects took the Duolingo test twice and their scores were correlated using Pearson correlation. For the third research question, equipercentile linking was performed to generate a concordance table that maps Duolingo test scores to TOEFL iBT total scores. Equipercentile linking is a preferred linking method to compare two different tests administered on the same subjects (ETS, 2010). Descriptive statistics and correlational analysis were conducted using SPSS 21. The R package EQUATE (Albano, 2014) was used to conduct equipercentile linking with log linear pre-smoothing (Kolen & Brennan, 2004).

3. Results

3.1 Descriptive Statistics of the Sample

Table 2 presents the mean, standard deviation, and range of the Duolingo English test scores, the total and subsection scores of the TOEFL iBT. To evaluate the performance of subjects in this research sample, the mean and standard deviation of the population of TOEFL scores in 2013 was obtained from the TOEFL website and presented in Table 2. For TOEFL total scores, the study sample has a 14.5 point higher mean than that of the 2013 population, and the standard deviation almost half of that of the 2013 population. This suggests that the study sample is more homogeneous with higher TOEFL scores. This is not surprising with the majority of subjects being college students.

Table 2. Means and Standard Deviations

Test	N	Mean	Standard Deviation	Score Range		TOEFL 2013 Population Mean	TOEFL 2013 Population Standard Deviation
Duolingo Test - First	214	6.91	1.82	.8-10			
Duolingo Test - Second	107	7.14	1.91	.5-10			
TOEFL Total	214	95.48	11.51	62-117		81.0	20
TOEFL Reading	214	25.42	3.75	13-30		20.1	6.7
TOEFL Listening	214	24.4	4.02	10-30		19.7	6.7
TOEFL Speaking	214	22.08	3.22	14-29		20.1	4.6
TOEFL Writing	214	23.69	3.58	12-30		20.6	5.0

Among these 214 subjects, 165 subjects uploaded TOEFL score reports and 49 subjects had self-reported TOEFL scores. These two groups of subjects were compared on their mean and standard deviation of TOEFL scores in Table 3, and on the correlation between TOEFL total scores and Duolingo scores. The self-reported scores have significantly lower means on the reading section than official scores, while there is no significant difference for total

scores or the other section scores. A z-test of difference in correlation coefficients from two independent samples indicates that the correlation between TOEFL total scores and Duolingo scores was not significantly different for the self-reported TOEFL scores (r = .70) and officially reported TOEFL scores (r = .62), z = .85, p > .05. Indeed, the correlation between the self-reported scores and the official score reports was nearly 1.0 for the 165 subjects who turned in the report. From these results, it was concluded that the self-reported scores are in general accurate, which is in agreement with earlier research on self-reported data (Pearson, 2009).

Table 3. Comparison of subjects with and without TOEFL score report

TOEFL iBT scores	Self-reported Data (N=49)		Official score report (N=165)		Mean comparison
	Mean	SD	Mean	SD	
Total	92.37	13.98	96.40	10.54	$t(65)=-1.87^{a}$
Reading	23.76	4.41	25.92	3.38	$t(65)=-3.16^{*}$
Listening	23.76	4.61	24.59	3.82	$t(69)=-1.15^{a}$
Speaking	22.08	3.55	22.07	3.13	$t(212)=.02^{a}$
Writing	22.78	4.39	23.96	3.26	$t(65)=-1.75^{a}$

Note. *p<.01. [a] t-statistics are based on results with unequal variances.

3.2 External structure evidence for validity

External structure evidence for validity refers to the relationships of assessment results to the results of other variables. One kind of external structure evidence answers the question, "How well does performance on this assessment procedure predict current or future performance on other criteria?" This is also known as criterion-related validity. To evaluate the external structure evidence of the Duolingo English test scores, the Pearson correlation coefficients between the Duolingo scores and the total and section scores of the TOEFL iBT were calculated and presented in Table 4.

A Pearson correlation coefficient measures the direction and the strength of the linear relationship between two variables. It ranges from -1 to +1. The sign of the coefficient indicates the direction of the relationship and the magnitude indicates the strength of the relationship with 0 indicating absence of linear relationship and 1 indicating a perfect linear relationship. The magnitude of the Pearson correlation coefficient depends on the restriction of range which happens when the test score has a narrow range or when the examinee group is homogeneous in ability so that the test scores do not vary across the whole score range. Restriction of range will reduce the magnitude of correlation coefficients. Based on the sample TOEFL scores, we can see that the current study sample is more homogeneous when compared to the 2013 population TOEFL takers. Thus, the correlation coefficients calculated as validity and reliability coefficients could be underestimated.

As the total score is the sum of all the section scores and more reliable, the Duolingo test scores are expected to have a higher correlation with the TOEFL total scores than the TOEFL section scores. The correlation with the TOEFL total score is substantial, while the correlations with the TOEFL section scores are moderate. The correlation with the TOEFL section scores is highest for speaking and writing, and lowest for reading.

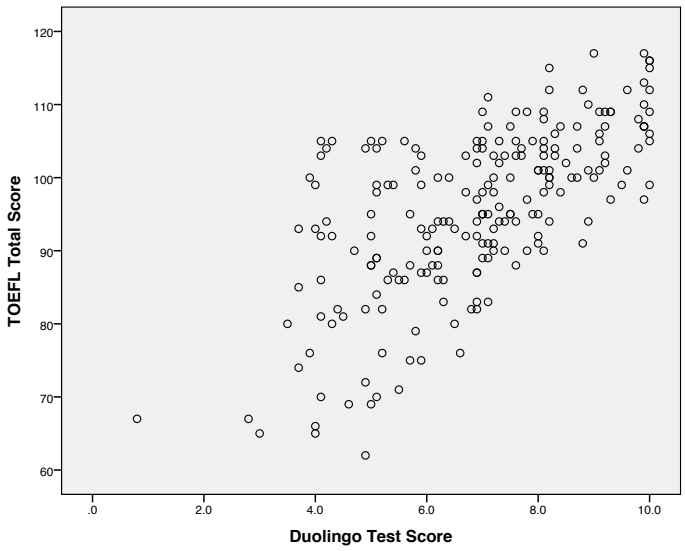
Table 4. Correlations between Duolingo and TOEFL scores (N = 214)

TOEFL score	Correlation
Total	.67
Reading	.45
Listening	.52
Speaking	.56
Writing	.56

Note: All correlations are significant at $p < .001$.

Figure 1 presents the scatterplot of TOEFL total scores and the Duolingo test scores. One subject had relatively low scores on both tests. For the other subjects, the TOEFL total scores are at the upper half of the scale, ranging from 62 to 117, while the Duolingo scores are more spread out along the scale. Note that for the subjects who scored relatively low on the Duolingo test, some of the TOEFL scores are at a higher range. For example, for Duolingo scores lower than 5, the corresponding TOEFL scores ranged from 62 to 105 with a mean of 85 and standard deviation of 13. However, for the subjects who scored relatively low on the TOEFL, their scores on the Duolingo test are also in the lower range. For example, for TOEFL scores lower than 80, the corresponding Duolingo scores ranged from .8 to 6.6, with a mean of 4.5 and standard deviation of 1.33. This suggests that the TOEFL scores tended to be the more "lenient" of the two.

Figure 1. Scatterplot of TOEFL Total scores and Duolingo test scores



3.3 Test-retest reliability

To evaluate the reliability of the Duolingo test scores, test-retest reliability coefficient was calculated for subjects who retested after two weeks by correlating the scores from the first and second Duolingo tests. Note that the Duolingo test is computer adaptive, which means that the same subject would most likely have been exposed to different items in the two test occasions. Thus, this reliability coefficient can also be considered as alternate form test-retest reliability. Although the use of different test items eliminates the effects of subjects' remembering specific items, it does not eliminate general practice effects.

For the 107 subjects who took the Duolingo test twice, the estimated reliability coefficient for Duolingo test scores was .79, indicating that the Duolingo test scores are reliable. Typically, the test-retest reliability coefficients for standardized achievement and aptitude tests with identical forms are between .80 and .90 (Nitko & Brookhart, 2011). However, test-retest reliability coefficients tend to be lower for tests of different items than for tests with identical forms. Test-retest reliability with different items is affected by measurement errors due to sampling different test items in addition to the measurement errors associated with the test-retest reliability with identical forms.

3.4 Equipercentile linking results

This section provides the equipercentile linking results for Duolingo test scores and TOEFL iBT total scores in a tabular form. Although very high correlations (> .86) are recommended for scores of different tests to be linked for concordance purposes (Dorans, 2004), moderately high correlations (.6-.8) in linking studies for concordance purposes are quite common (e.g., Cartwright, 2012; ETS, 2010; Northwestern Evaluation Association, 2011; Pearson, 2009). The concordance table after equipercentile linking will list the TOEFL total score for each score point on the Duolingo test. As equipercentile linking defines concordance between score scales by setting equal the cumulative distribution functions for the two scales, the concordance is nonlinear. For example, the mid-point of the Duolingo test scale is 5.0, corresponding to 82 on the TOEFL scale. A Duolingo score of 7.2 corresponds to the TOEFL score of 100, while a Duolingo score of 9.4 corresponds to the TOEFL score of 115. The concordance table only suggests that the corresponding scores have similar percentile rank in this sample distribution. It does not necessarily mean that the scores from the two tests are interchangeable.

When the sample is large enough and with more cases in each score band, the results from linking would be more accurate. However, as the current study has 214 subjects, there are only a few scores for each of the 100 score points from 0.0 to 10.0. Thus, instead of reporting the concordance table for each score point on the Duolingo test, Table 5 presents concordance results in intervals of .5 on the Duolingo scale. The shaded rows indicated that the sample sizes within these score bands are greater than 5% of the total sample size, and these results are more reliable. For the rest of the score bands in the lower end of the range, caution should be used when interpreting the results.

Table 5. Concordance between TOEFL total score and Duolingo test scores

Duolingo	TOEFL
9.6-10.0	117-120
9.1-9.5	114-117
8.6-9.0	110-113
8.1-8.5	107-110
7.6-8.0	103-106
7.1-7.5	99-102
6.6-7.0	95-98
6.1-6.5	91-95
5.6-6.0	87-91
5.1-5.5	83-87
4.6-5.0	79-82
4.1-4.5	74-78
3.6-4.0	69-73
3.1-3.5	64-68
2.6-3.0	58-63
2.1-2.5	52-57
1.6-2.0	44-50
1.1-1.5	35-42
0.6-1.0	23-33
0.0-0.5	2-20

4. Conclusion

This study assesses some of the evidence for validity and reliability of the Duolingo English test for non-native English learners. In addition, the Duolingo test scores were linked to TOEFL iBT scores to establish concordance. Scores from the Duolingo English test were found to be substantially correlated with the TOEFL iBT total scores, and moderately correlated with the individual TOEFL iBT section scores, which present strong criterion-related evidence for validity. The Duolingo test scores present high test-retest reliability over a two-week interval. Equipercentile linking was used to establish concordance between TOEFL scores and the Duolingo test scores. Duolingo English test scores are on a scale of 0-10 and TOEFL iBT total scores are on a scale of 0-120. For international students to apply for studying in US universities, the minimum cut-off score of TOEFL iBT is 80 and a more selective cut-off score is 100, corresponding to scores 5.0 and 7.2 respectively on the Duolingo English test.

There are some limitations to this study. First, the preliminary administration of the Duolingo English test for this study was not high-stakes, so the effort that subjects put into completing the test are expected to be lower than when they took TOEFL. This may have an impact on the results, and may partially explain why Duolingo test scores were spread out almost the entire scale. Second, the results of this study were sample dependent. As the current study had a more homogeneous sample with higher English proficiency than the general TOEFL takers and Duolingo users, the concordance at the lower end of the score range may be inaccurate. Third, the sample size is relatively low for linking studies to achieve the ideal precision for concordance purposes. Once the Duolingo English test is available to the public, future studies could be designed to administer the Duolingo test within a higher-stakes environment, with a larger pool of subjects to draw from.

References

Albano AD (2014). equate: Observed-Score Linking and Equating. R package version 2.0-2, URL http://CRAN.R-project.org/package=equate.

Cartwright, F. (2012). Linking the British Columbia English examination to the OECD combined reading scale. Retrieved May 8, 2014, from http://www.bced.gov.bc.ca/assessment/linking_bc_eng_pisa.pdf.

Dorans, N.J. (2004). Equating, concordance, and expectation. Applied Psychological Measurement, 28(4), 227-246.

Educational Testing Service. (2010). Linking TOEFL iBT scores to IELTS scores—A research report. Princeton, NJ: ETS.

Kolen, M. J., & Brennan, R. L. (2004). Test equating methods and practices. New York, NY: Springer-Verlag.

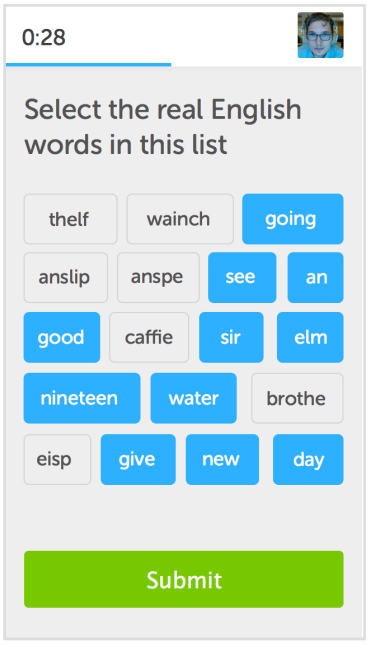
Nitko, A. J., & Brookhart, S. (2011). Educational Assessment of Students, 6th edition. Englewood Cliffs, NJ: Merrill. (ISBN: 0-13-097781-0)

Northwestern Evaluation Association. (2011). Wyoming linking study: A study of the alignment of the NWEA RIT Scale with the Proficiency Tests for Wyoming Students (PAWS). Retrieved May 8, 2014, from http://www.nwea.org/sites/www.nwea.org/files/resources/WY_Linking%20Study.pdf.

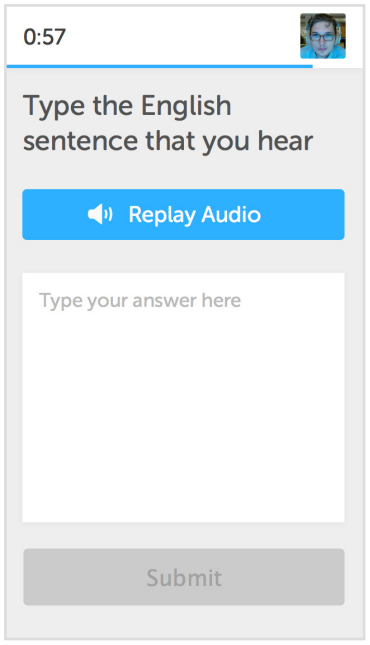
Pearson. (2009). Preliminary estimates of concordance between Pearson Test of English Academic and other Measures of English language competencies. Retrieved May 8, 2014, from http://www.pearsonpte.com/SiteCollectionDocuments/PremliminaryEstimatesofConcordanceUS.pdf.

Appendix 1: Item types of the Duolingo English test

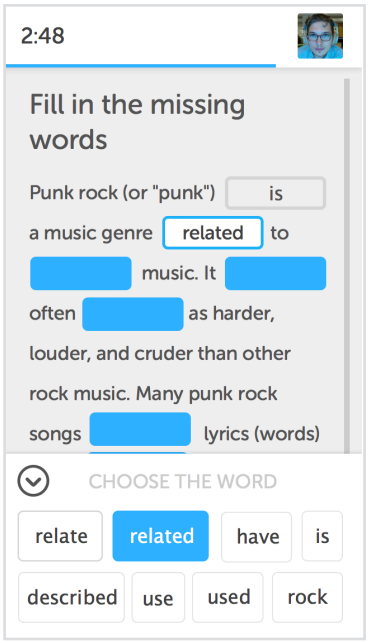
Vocabulary



Listening and Transcription



Sentence Completion



Speaking