The Reliability of Duolingo English Test Scores

Duolingo Research Report DRR-16-02
August 1, 2016 (5 pages)
englishtest.duolingo.com/resources

Burr Settles

Abstract

This study reports several reliability measures for the first operational year of the Duolingo English Test (DET). Our results show that during this time, the standard error of measurement (SEM) was stable across the score range, reliability and internal consistency coefficients were both 0.96, and the test-retest reliability coefficient was 0.84. Repeat test-takers exhibited some practice effects, improving their score by 1–3 points on average, but the overall effect size (0.10) was small. These findings indicate that DET scores are reliable.

Keywords

Duolingo English Test, reliability, internal consistency, repeater analysis

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

The Duolingo English Test (DET) is an efficient, affordable, and accessible option for English-language proficiency testing. The DET can be administered on-demand anywhere in the world, through an online software application called Duolingo Test Center, available on the Web[1] or mobile devices running Android or iOS. It was launched "in beta" during July 2014, and after additional research and development was deployed to all device platforms by June 2015.

The DET is a computer-adaptive test of general English-language ability that costs about US$50. Tests usually last 10–25 minutes, and includes interactive test item formats that take advantage of the modern speech technologies, camera, microphone, and other sensors available on even the least expensive computers, phones, and tablets around the world today. Each DET test is reviewed by a team of remote human proctors using the device's built-in camera and microphone, and results are reported in less than 48 hours (median 18.5 hours). Scores are reported on a 100-point scale.

An important measure of quality for any assessment is the *reliability* of its scores, which refers to the overall consistency of the test. To be considered reliable, a test should produce similar results under similar conditions, such as if a test is repeated or administered using different test items to the same group of people. However, language testing — as with any kind of measurement — can be influenced by a number of factors that are not relevant to the examinee's actual language proficiency. These irrelevant factors contribute to what is called "measurement error," which in turn determines the reliability of test scores. A well-designed test should yield a score that reflects a person's true ability as much as possible and keep measurement error to a minimum.

Reliability can be quantified using a variety of statistical indexes designed to evaluate the consistency of test scores. In this study, we report on several reliability estimates for the first full year of operational DET tests. In particular, this report focuses on:

- *Reliability based on standard error of measurement (SEM)*, which quantifies the measurement error,
- *Internal consistency reliability*, or how comparable results are among different test items, and
- *Test-retest reliability*, or how consistent scores are for "repeaters" who took the test multiple times.

2 Data source

Our data consist of scores for all operational DET tests administered from July 1, 2015 through June 30, 2016. This is because the test was deployed to different platforms at different times (Web, Android, and iOS), and because several institutions began accepting DET scores to satisfy English-language requirements in Summer 2015[2]. We believe this first full year of high-stakes use, with availability on all three software platforms, best represents the performance of a stable testing population. Note that we only consider valid test results in this study; that is, we omit any test scores where the candidate was caught cheating, experienced technical difficulties, or where the score could otherwise not be computed, verified, and reported.

In September 2015, Duolingo began offering a "practice" test in addition to the paid and proctored "certified" test. The practice test is free and requires no proctoring or security protocol. But it draws from a smaller pool of test items that are less discriminative than the certified test, has less rigorous stopping criteria (median test length is 8 minutes for the practice test and 17 minutes for the certified test), and does not provide a certified score. For comparison, we provide reliability estimates for both kinds of tests.

For test-retest reliability analyses, we defined a "repeater" to be anyone who took a DET test, and then took another DET test

Corresponding author:
Burr Settles, Staff Research Scientist & Engineer
Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA

of the same type within 30 days. The rationale for this selection method is that significant language learning is unlikely to occur over such a short period of time. Presumably, the score changes between tests can be mostly attributed to measurement error and the quality of test scores. During the one-year time-frame of our analysis, about 16% of the examinees repeated the certified test at least once, and about 3% took the test three or more times.

Table 1. Sample sizes for the data in this report.

Test Type	All Tests	Repeaters
Certified	8,130	1,647
Practice	1,307,989	413,273

Table 1 summarizes our data sample sizes. The certified test examinees represent 128 countries, and the repeaters represent 83 different countries. Outliers (i.e., test-takers with extremely large score increases or decreases) were omitted from our test-retest analyses, since these extreme score changes were most likely due to spurious factors other than proficiency or measurement error. We did this by excluding test pairs whose score change was greater than ± 87 (three standard deviations of all operational test scores). Out of the 1,657 tests that were repeated within 30 days, 10 of them (0.6%) were removed from further analysis.

3 Results

3.1 SEM-based reliability

In educational measurement, reliability is usually expressed by a *reliability coefficient*, which is a statistical index ranging from 0 (not at all reliable) to 1 (perfectly reliable). This coefficient can be estimated in a variety of ways depending on the intended use and underlying theory. Since a person's "true ability" score is a theoretical construct and can therefore never be obtained directly, reliability coefficients are estimated using statistical methods. One interpretation of a reliability coefficient is the correlation between a real test score result and the test-taker's theoretical true ability score.

Test score precision can also be expressed by the *standard error of measurement (SEM)*, which defines a score range within which the true ability score probably lies. SEM has a complementary relationship with reliability: the smaller the SEM, the higher the reliability coefficient. Classical SEM depicts the amount of measurement error for a typical (average) test-taker. However, according to the *Standards for Educational and Psychological Testing* [1], test publishers should instead report *conditional SEM (CSEM)* estimates, which depict the amount of measurement error for each test score along the ability scale.

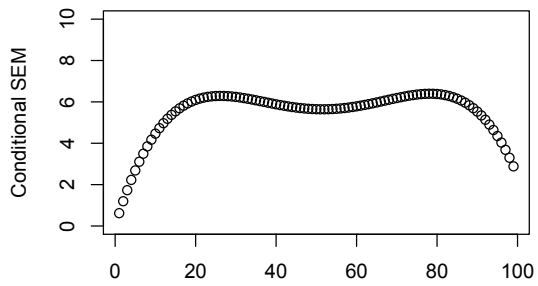

Figure 1. Conditional SEM for certified test scores.

Figure 1 shows the conditional SEM results for the complete DET score range (0–100) of the certified test. These were computed using the Mollenkopf-Feldt method, also called the polynomial method [2, 3], which is recommended for its ability to smooth out the fluctuations in CSEM point-estimates that result from varying numbers of test-takers at each score point [3]. Error is highest in the middle of the score range, where it is very stable between 20 and 85, and lower at the two extremes (which is common). CSEM is reported in the same units as the test score and represents one standard deviation. In other words, if a person scores 56 on the DET, which is roughly the minimum score for English-language university admissions, there is a 68% chance that her true ability score lies in the range 56 ± 5.5 $(51, 62)$.

We estimate the overall SEM and reliability coefficient using the following standard formulas [4]:

$$\text{SEM} = \sqrt{\mathbb{E}[\text{CSEM}^2]}\,, \quad \text{reliability} = 1 - \left(\frac{\text{SEM}}{\sigma}\right)^2\,,$$

where $\mathbb{E}[\cdot]$ denotes the expected (mean) value across the score range, and σ is the standard deviation of all test scores.

Table 2. Score summary, standard error of measurement (SEM), and reliability coefficient for operational tests.

Test Type	μ	σ	SEM	Reliability
Certified	56.3	29.1	5.5	0.96
Practice	31.5	21.7	9.8	0.80

Table 2 reports the mean (μ) and standard deviation (σ) of the overall score populations, along with SEM and derived reliability estimates, for both certified and practice tests. The first observation is that certified test scores are, on average, about 25 points higher than practice test scores (albeit with a smaller σ, possibly due to a more restricted range). This is expected, since the certified test is mainly taken for high-stakes purposes, while the practice test can be taken for free by anyone for any reason, and probably attracts a lower-proficiency audience. SEM and reliability are also better for the certified test: $5.8 < 9.6$ and $0.96 > 0.80$, respectively. This is likely due to the more stringent testing conditions, item requirements, and length of the certified test.

Importantly, the certified test reliability coefficient is well above 0.9, which is considered to be the minimum threshold for tests "intended for individual diagnostic, employment, academic placement, or other important purposes" [5]. The results show that DET scores satisfy this criterion.

3.2 Internal consistency reliability

For internal consistency reliability — that is, the degree to which different test items produce similar results — the most common measure is Cronbach's α [6]. This statistical index represents an aggregation of pairwise correlations between all test item scores. However, that calculation requires identical test forms for all individuals, and it is not robust to sparse or missing data. So it cannot be used to evaluate computer-adaptive tests like the DET, since each test administration contains different items drawn from a large pool.

Instead, computer-adaptive test developers often use the *split-half method* for internal consistency [7]. This involves splitting the test item bank into two equal halves, and then scoring each

test twice using the items from each half independently. The split-half reliability coefficient is simply the correlation between the test scores from these two halves.

Interestingly, Cronbach's α is equivalent to the mean of all possible split-half coefficients of a test [8]. So Cronbach's α and the split-half method represent two extremes on a continuum of test granularity for estimating internal consistency. Drawing inspiration from the "hashing trick" in machine learning research [9], which solves an analogous problem in large sparse data sets by representing middle-ground between these extremes, we propose and report on a novel internal consistency measure called the *hashing-α method*. This works by splitting test items into k distinct partitions (by applying a hashing function to each item ID, modulo k) and then scoring each test k times using the items from each partition independently. This is very similar to the split-half method if $k = 2$. The hashing-α coefficient is calculated using Cronbach's α, as though each of the k partition scores were test item scores in and of themselves.

Table 3. Internal consistency reliability coefficients.

Test Type	Split-Half	Hashing-α
Certified	0.96	0.93
Practice	0.83	0.81

Table 3 reports internal consistency results for certified and practice tests using both methods. We used $k = 20$ for the hashing-α method, since there are approximately 20 items per certified test[3]. Not surprisingly, reliability for the certified test is again higher than the practice test. Both internal consistency measures are well above 0.9 for the certified test, again indicating that it is very reliable.

3.3 Test-retest reliability

As with any large-scale assessment, it is not uncommon for some DET examinees, for any number of reasons, to take the test more than once. It is also not uncommon for such "repeaters" to obtain different scores for each test session, even if the sessions were administered relatively close together in time. This is again due to measurement error (which can be further compounded by multiple sessions). However, a valid and reliable test should produce scores that only marginally vary from one repeater's session to the next.

The *test-retest reliability* coefficient — measured as the correlation between a repeater's two consecutive test scores — is one of several measures of interest regarding empirical score variation among repeaters. In this section, we report on some demographic characteristics of DET repeaters, how they performed on their first and second tests, the test-retest reliability coefficient, and the effect size of any differences between their two scores.

Table 4 summarizes the DET certified test repeaters by country, as determined by their Internet connection. Most of the repeaters in this study were located in Asia or the Americas. For comparison, the last column shows the percentage of all tests that were administered per country. China, Italy, and Japan seem most over-represented among repeaters, with a more than 3:2 ratio of repeaters relative to the overall population. The vast majority of repeaters (90.4%) took both tests using the same platform (Web, Android, or iOS). Of those who repeated using a different platform, half of them (5.1% of repeaters) took the first test on Web and the second test using a mobile device, while about a quarter (2.7% of repeaters) did the opposite. Neither country nor test platform showed any significant effects on test score changes.

Table 4. Physical location of repeaters (vs. all tests).

Country	Count	Percent	All Tests
China	424	25.6%	16.0%
United States	257	15.5%	11.5%
Brazil	247	14.9%	17.8%
Mexico	205	12.4%	11.8%
Columbia	76	4.6%	5.6%
Italy	60	3.6%	2.3%
India	44	2.7%	2.9%
Guatemala	36	2.2%	1.9%
Canada	25	1.5%	1.8%
Japan	23	1.4%	0.9%

Note: $N = 1{,}647$ certified repeaters; showing top 10 countries.

Table 5 summarizes test scores and score changes for repeaters, as well as test scores for the overall population of operational tests (μ, σ, and various percentiles). The repeaters' performance on their first test was lower than that of the overall operational group; this can be seen for both the mean score and at all percentile levels. On their second test, repeaters marginally improved their scores, but were still no higher than the overall group on average. This trend of below-average candidates slightly improving their score is consistent with studies of TOEFL® iBT test repeaters [10].

The test-retest reliability coefficient was 0.84, which is moderate to high for a computer-adaptive test. Such coefficients typically range from 0.8 to 0.9 for standardized tests with identical forms [11], but computer-adaptive tests are often less reliable because the items used will vary from one test session to the next. Even so, DET reliability is comfortably within this recommended range for standardized tests. Note that a previous independent study reported the DET test-retest reliability coefficient to be 0.79 [12]. This in fact still the reliability for the practice test. However, subjects in the previous study were recruits[4] and therefore did not represent an organic population of DET examinees or repeaters. They also used a preliminary version of the DET, before security protocols were implemented and before it was made publicly available. The current study represents a more stable and realistic population of examinees using the operational DET, with an item bank that has been refined over time.

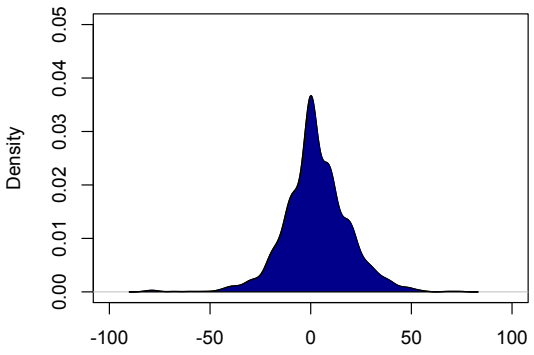

Figure 2. Distribution of repeater score changes.

Table 5. Statistics for certified test repeater scores and score changes, reliability, effect size, and all operational test scores.

	μ	σ	1%	25%	50%	75%	99%	Reliability	Effect Size
First test	52.4	27.6	0	31	58	71	99	0.84	
Second test	55.2	28.7	0	32	60	80	99		
Score change	+2.8	15.9	-40	-6	+1	+11	+42		0.10
All operational tests	56.3	29.1	0	32	60	81	99		

Note: $N = 1{,}647$ repeaters; $N = 8{,}130$ total operational tests.

Figure 2 depicts the distribution of score changes for certified tests. This distribution appears approximately symmetrical and normally-distributed (bell-shaped) around zero. This is consistent with the percentile results in Table 5, and suggests that the majority of repeaters experience a small score change, ranging from zero to a few points in either direction. The average score change went up by 1–3 points, which suggests a small practice effect (i.e., gain due to familiarity with the test instrument). To assess the significance of these gains, we also report the *effect size*, also known as Cohen's d, which is a standard scale-free measure for the relative size of the score change between two tests [13]:

$$\text{effect size} = \frac{\mu_2 - \mu_1}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}} .$$

This is essentially the difference between two test score means divided by their pooled standard deviation. The effect size is only 0.10, which according to Cohen's rule [13] is less than 0.20 and should be considered small.

Qualitatively, we also investigated the 16 repeaters with a score change in the 99th percentile. Since the DET includes a video transcript of each test session in its entirety — the primary purpose of this for proctors to verify the identity and integrity of each examinee — we were able to subjectively assess the proficiency of these extreme candidates by reviewing their videos. In all cases, we thought the repeater deserved the second, higher score, but for whatever reason performed extremely poorly on their first attempt.

4 Conclusions

Test score reliability is an important property of any assessment tool. In this report, we examined several reliability measures using data from the first full year of operation for the Duolingo English Test (DET). Results show that for the certified DET test, the standard error of measurement (SEM) is stable across the score range, SEM-based reliability and internal consistency reliability coefficients are both 0.96, and the test-retest reliability coefficient is 0.84 with only a small effect size in score change. These results suggest that the certified DET scores are very reliable.

It is important to note that test *reliability* (the extent to which test scores are consistent) is different from test *validity* (the extent to which test scores measure what they claim to measure). Previous research has found DET scores to be significantly correlated with TOEFL iBT scores [12], IELTS scores [14], and university faculty evaluations of international students in the United States [15]. Those results provide criterion validity evidence for DET score use. Taken together, the research to date indicates that certified DET scores are not only convenient for test-takers and other stakeholders to obtain, but they are also valid, reliable, and appropriate as a high-stakes English-language assessment.

Acknowledgements

The author would like to thank Feifei Ye and Klinton Bicknell for research pointers and statistical advice during the preparation of this report. Thanks also Eleanor Avrunin, Jefferey Brenzel, Masato Hagiwara, and Bożena Pająk for additional comments that improved the final manuscript.

Notes

1. https://englishtest.duolingo.com
2. Examples include the Harvard Extension School in the United States, Ashton College in Canada, and the Max Planck Institute in Germany.
3. Note that this yields a pessimistic estimate, since most of the k partition scores are based on a single test item or no items at all (thus falling back to the Bayesian prior used by the computer-adaptive algorithm).
4. The previous study mostly consisted of international students recruited from American universities, so their test scores were significantly higher than the candidates from the present study. This suggests a "truncated sample" bias, which can also artificially deflate the reliability coefficient.

References

[1] AERA, APA, and NCME. *Standards for Educational and Psychological Testing*. AERA, 2014.

[2] W.G. Mollenkopf. Variation of the standard error of measurement. *Psychometrika*, 14:189–229, 1949.

[3] L.S. Feldt, M. Steffen, and N.C. Gupta. A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, 9(4):351–361, 1985.

[4] N.S. Raju, L.R. Price, T.C. Oshima, and M.L. Nering. Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, 31(3):169–180, 2007.

[5] R.F. DeVellis. *Scale Development: Theory and Applications*. Number 26 in Applied Social Research Methods. SAGE Publications, 2011.

[6] L.J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[7] K.R. Murphy and C.O. Davidshofer. *Psychological Testing: Principles and Applications*. Pearson, 2004.

[8] N.J. Salkind. *Encyclopedia of research design*, volume 1. SAGE Publications, 2010.

[9] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1113–1120. ACM, 2009.

[10] Y. Zhang. Repeater analyses for TOEFL iBT. Technical Report RM-08-05, ETS, 2008. http://bit.ly/2aqQQKq.

[11] A.J. Nitko and S. Brookhart. *Educational Assessment of Students*. Merrill, Englewood Cliffs, NJ, 2011.

[12] F. Ye. Validity, reliability, and concordance of the Duolingo English Test. Technical report, University of Pittsburgh, May 2014. http://bit.ly/1PCG9Ba.

[13] J. Cohen. *Statistical Power Analysis for the Behavior Sciences*. Lawrance Eribaum, Mahwah, NJ, 1988.

[14] M. Bézy and B. Settles. The Duolingo English Test and East Africa: Preliminary linking results with IELTS and CEFR. Technical Report DRR-15-01, Duolingo, 2015. http://bit.ly/1SJjrHk.

[15] L. Ishikawa, K. Hall, and B. Settles. The Duolingo English Test and academic English. Technical Report DRR-16-01, Duolingo, 2016. http://bit.ly/29WsDcu.