【資料】

Exploring the Duolingo English Proficiency Test at the BECC

Kelly Eileen Rose

BECC における Duolingo 英語検定試験の研究

ケリー・アイリーン・ローズ

Abstract

This paper describes a study examining the effects of the computer-adaptive Duolingo English Test (DET) among a small group of first and second year students in the Global Communication Department at Hiroshima Bunkyo Women's University. This test provides a specific DET score and correlates to the widely utilized English language proficiency tests Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS). In addition, the DET score is aligned with the Common European Framework of Reference (CEFR). Therefore, by undergoing the DET, an examinee can gauge her English language level in a variety of milieu. In this project, qualitative data in the form of interviews and anonymous surveys as well as the quantitative data of the participants' DET scores are examined in order to ascertain the participants' English proficiency levels, as well as their motivation and confidence levels for studying English. The results indicate that utilizing this DET assessment can positively affect participants' motivation for studying English and can be a beneficial tool for tracking English language progress.

Introduction

With the rapid progression of technology and its integration into our daily lives, educators should actively explore new advancements and how they can affect our students. This study examines how an English test administered entirely on a computer, using an unfamiliar format, affects English language learners. Can such a novel proficiency test become a widely used method of gauging English language ability? This study trials the cutting-edge Duolingo English Test with English language learners in their first and second years of university, and examines its potential as an additional English language testing method used at Hiroshima Bunkyo Women's University Bunkyo English Communication Center (BECC).

The Duolingo English Test

In 2014 the popular language learning software company, Duolingo, launched its computer-

adaptive English proficiency test– the Duolingo English Test (herein DET). Originally, it was available on a variety of platforms and administered/proctored completely online allowing participants living anywhere with a stable internet connection an opportunity to take the examination. Priced at $49/5900 yen it is significantly less than its TOEFL/IELTS counterparts. For test security, participants must be in a quiet, private space and show a government-issued photo ID with each test monitored remotely by a human. It tests English knowledge across the major language skills and uses a web camera and microphone to record participants' speaking portions of the test. The turnaround time for grading is within two days of taking the test with scores being emailed to participants, and provided to participating universities free of charge. In addition, all speaking and writing portions of the test are recorded and can be viewed by admissions personnel. Not only does this unique feature allow admissions staff a better idea of each candidate's English abilities, but it has the potential to serve as a way for participants to gain valuable feedback for future improvement.

Once scores are compiled and released, participants can know not only their specific DET score, but also their Common European Frame of Reference (CEFR) level. Originally designed by the Council of Europe, the CEFR is well-established within the global linguistics community as a useful tool for categorizing language abilities within the reading, writing, listening, spoken production and spoken interaction language skills. The DET is also highly correlated with the TOEFL and the IELTS. This allows a participant to take the DET for a more reasonable price and in a more convenient setting and still have a good frame of reference as to what their level might be when taking other, more expensive proficiency tests. Moreover, the DET is growing in popularity among English-speaking universities. Students wishing to study abroad at such institutions can submit DET scores directly to participating universities and these results can serve as an additional supplement to TOFEL or IELTs scores. In 2017, responding to feedback from university administrators the DET added an essay writing component to the test and included all writing as well as speaking portions of the test to be viewed after the test by admissions personnel.

One key feature of the DET is that it is computer adaptive, meaning the program changes each test to suit the specific level of its examinee. As one of the lead DET researchers, Jeffery Brezel explains, "...we use machine learning and natural language processing technology to fit statistical models to tens of thousands of CEFR-annotated text passages and wordlists...In this way the DET is capable of generating and calibrating a large number of test items, while still achieving inter-item and test-retest reliability metrics that exceed industry standards" (Brenzel & Settles, 2017) In addition, test security is strengthened since it is nearly impossible for a test to be duplicated. This benefits test-takers by providing a customized test suited to each person's English abilities in much less time than a traditional paper-based test (Ishikawa, Hall, & Settles, 2016) .

The DET items themselves also do not resemble those found in a traditional standardized test, but instead blend language skills and cognitive processes. Brenzel describes one portion of the test, "...requires examinees to discriminate between real English words (e.g., *meeting* and

thunder) and English-like pseudowords (e.g., *clerm* and *earts*). Though this is not obviously related to everyday language tasks, it does require the same cognitive processes (e.g., lexical and morphological activation) that are used in everyday reading, writing, and even listening activities" (Brenzel & Settles, 2017). In both the speaking and writing portions, examinees must respond to several prompts, with some containing photographs. There is also a newly-added extended writing portion requiring examinees to respond to a prompt aiming for 50−100 words. Additionally, each portion of the DET is timed with listenings restricted to just three instances. By drawing upon a wide corpus of CEFR-aligned texts, and using interactive items which measure several language skills at once, an accurate assessment can be generated in less time than other standardized proficiency tests.

Student Background

This study involved a small group of twenty-four university students in a private women's university in Hiroshima, Japan. The first Phase in January 2017 included nine students at the end of their first year. The second Phase in December 2017 included a total of 19 students. Of the second Phase group fifteen were first year students and four were returners from Phase (at the end of their second year). All students were part of the Global Communication Department (GCD) which involves a great amount of English language instruction–especially within the first two years. The majority of this English instruction occurs within the Bunkyo English Communication Center (BECC) and the Self-Access Learning Center (SALC)–two English-speaking facilities within the university. A compulsory part for students in this department is to frequently take the Test of English for International Communication (TOEIC) which is a widely used business English assessment in Japan. Students are not required to undergo any other testing such as TOEFL or IELTS, however some voluntarily do for study abroad preparation. Prior to this study all of the participants had no experience with Duolingo's English Test, however they were familiar with its language learning app, as it was included in their first year English Communication course.

Methodology

The study's duration spans the 2017 year with Phase 1 being administered in late January and Phase 2 in late December. With the first instance, nine participants at the end of their first year in the Global Communication Department took the DET in the SALC multi-purpose rooms. These were small, enclosed booths within the facility with an English conversation lounge located just outside the rooms. The test was administered on an ipad mini using the Duolingo English Test app. A week prior to the test participants were given information about the test procedures and asked to familiarize themselves by attempting a practice test, available in the DET app. Immediately before the test the author verbally discussed the test procedures to prevent tests being voided by DET security. The test itself took between twenty to thirty minutes for participants to complete. After the test, participants engaged in a group interview about their impressions of the test. Upon receiving their test results students took an anonymous survey in English and

Japanese regarding their impressions of the DET. Both of these documents can be viewed in the Appendix.

The second iteration of the project occurred in December 2017 and included nineteen participants. Of the Phase 1 group in January, four students participated again, having completed nearly two years in the Global Communication Department. The remaining fifteen were first year GCD students. The testing took place mostly in the SALC multi-purpose rooms with the exception of two tests being administered within empty classrooms. Unlike Phase 1, the DET was administered entirely on a laptop computer using google chrome's Internet browser instead of the DET app on participants' iPad minis. Prior to the test participants were provided with background information on the test and procedures, however they could not try the practice test because they lacked the required hardware. Therefore, on the day of the test participants were verbally instructed about the procedures by the author, then took the DET practice test immediately before taking the real test. Similar to Phase 1 brief interviews were conducted immediately after the test and a survey of participant impressions was given after receiving results.

Variations Between the Two Phases

The required technology and access to it differed greatly between the two Phases. When the DET was first attempted in January 2017 it was available as an app to download onto students' iPad minis. Participants could attempt the practice test included in the app at their leisure and familiarize themselves prior to taking the DET. Participants used their iPad minis to record all portions of the test, including the writing, although the longer writing passage was not included yet. Some participants used a separate Bluetooth-connected keyboard but most used the touch keyboard standard to mobile devices. Between the first and second Phases of the project the DET underwent several changes. Most notably, Duolingo discontinued access to mobile devices, citing efforts to provide a more consistent experience for its users (Brenzel, Settles & Tousignant, 2017). Unfortunately, all this study's participants lacked access to a computer with a web camera/ microphone in a private environment. This required the author to procure acceptable laptops and install the necessary software. It also prevented participants from trying the DET practice test until just before attempting the test. This increased the total testing time and mental burden on participants.

The testing environment, specifically relating to noise level, also varied between the two Phases. The first Phase of the project occurred in late January during the university's examination week. During this time the SALC was quiet and calm with hardly any students visiting the conversation area nearby. However, the second Phase in December was a much busier time in the SALC and DET testing often occurred at times highly-frequented by students and teachers. This reportedly affected students' concentration, as indicated in qualitative data, with some students even having their tests disqualified by DET proctors due to excessive background noise.

It was also noticed that test security seemed to be stricter during the second Phase of the

project. Of the 19 participants in Phase 2, six had their tests rejected and another three had to resubmit their identification information. This contrasted with Phase 1 where none of the first group had any of these issues.

Findings

The majority of the Phase 2 participants in both cohorts (first and second year students) fall within Duolingo's "intermediate" category of 36%−55% within the B1 CEFR band. The average score of the Phase 2 for both cohorts is 46% correlating to a TOEFL score of 78, IELTS 5.5 and a B1 CEFR level. Of the Phase 1 group just over half are in the DET's "elementary" category ranking in the A1-A2 CEFR range with the others in the "intermediate" category within the B1 CEFR band. The average score for Phase 1 is 34% correlating to a TOEFL score of 67, IELTS 5 and a B1 CEFR level. For detailed category descriptors and correlated TOEFL/IELTS scores please see the Appendix.

Table 1 shows all participants with corresponding DET categories and CEFR levels.

DET Category / CEFR	Phase 1 JAN 2017 1st year students	Phase 2 DEC 2017 2nd year students	Phase 2 DEC 2017 1st year students
Elementary A1-A2	5	1	1
Intermediate B1	4	4	13
Advanced B2	0	0	1
Total Participants	9	4	15

Table 2 shows the Phase 2 first year participants' DET scores/categories with CEFR levels.

Student	DET Score	DET category	CEFR
A	57	advanced	B2
B	50	intermediate	B1
**C	35 / 34	elementary	A2
D	47	intermediate	B1
**E	50 / 54	intermediate	B1
**F	43 / 44	intermediate	B1
G	42	intermediate	B1
**H	54 / 50	intermediate	B1
I	36	intermediate	B1

*J	36	intermediate	B1
*K	52	intermediate	B1
L	52	intermediate	B1
M	48	intermediate	A2
*O	48	intermediate	B1
P	39	intermediate	B1

* These participants were required to resubmit their identification.

**These participants retook the DET due to their initial survey being disqualified by the DET reviewer. Only their final score was included in analysis (bolded).

Among the four students who took the DET for a second time in December 2017 all fall within the "intermediate" category. Of that cohort, two students showed growth (previously ranking in the "elementary" category), one stayed at about the same score, and the other decreased her DET score but remained within the "intermediate" category.

Table 3 shows the DET scores/categories and CEFR levels of the Phase 1 participants as first years and returning Phase 2 as second years.

Phase	Student	DET Score	DET category	CEFR
1	A	20	elementary	A1/ A2*
**2		54	intermediate	B1
1	B	26	elementary	A2
2		42	intermediate	B1
1	C	40	intermediate	B1
**2		42	intermediate	B1
1	D	50	intermediate	B1
2		44	intermediate	B1
1	E	28	elementary	A2
1	F	49	intermediate	B1
1	G	28	elementary	A2
1	H	41	intermediate	B1
1	I	20	elementary	A1/A2*

* The DET / CEFR correlation does not extend below a DET score of 26.

**These participants had their test disqualified by the DET reviewer. Both chose not to re-take their Phase 2 DET within this study's publication date.

The DET challenged the participants, as reported from their comments in the interviews and surveys. The most common adjective used during the interviews was "difficult." Most notably the speaking, writing and vocabulary portions of the test. However, this is to be expected by the computer-adaptive nature of the DET. Additionally, because the DET utilizes an untraditional item structure and each item has a fixed time limit, the testing experience was completely unfamiliar

to all of the first-year participants in both phases. Interestingly, students in both phases also indicated in the survey that they specifically liked the many speaking opportunities provided in the DET. Others liked the balance among language skills within the test with one commenting, "I can know my skill about speaking, writing, and listening." One student mentioned that she liked thinking and answering in "real-time." Finally, another enjoyed the short nature of the DET and that the time limits allowed her to progress through the test without worrying about her previous responses.

Despite the challenging nature of the DET all participants in both phases indicated during the interviews they would like to participate in this study if offered again in the future, and reiterated their willingness to use the DET again in the anonymous survey (see Table 4). Most comments from both surveys recognized the DET as a useful method for diagnosing strengths and weakness and tracking progress. One participant commented, "I think it is good test to understand my level." Another expressing her eagerness to repeat the DET, "I want to compare my score each year." Two mentioned specifically the DET's price stating, "It is cheaper than other English test" and "It's expensive for me to take all test, so maybe I continue to use this." The majority of participants also commented in both the interviews and surveys about their motivation for studying English increasing as a result of participating in this study. As one student wrote in the Phase 2 survey, "my motivation is up, I will study hard." It seems the DET's unconventional testing items paired with its computer-adjusted difficulty level made a positive impression upon participants. These comment shows us that despite its challenging design, the DET can be a useful tool for English language students at the BECC.

Table 4 shows the likelihood of participants willing to try the DET again in the future.

How likely are you to continue using the Duolingo English Test in the future?			
	Highly Likely	Somewhat likely	Not likely
Phase 1	3	5	0
Phase 2	6	7	1

Discussion

An area of interest is using the DET as a diagnostic tool particularly by providing test takers detailed feedback on their recorded speaking and written response portions. Currently the DET provides university admissions staff not only an examinee's DET score and correlating TOEFL/ITELS/CEFR levels, but a link to view the examinee's entire speaking and writing portions of the test. While intended by Duolingo to be used for study abroad candidates to offer a better picture of true English language abilities, it holds other benefits as well. In one manner, it could be used to improve an examinee's test-taking skills. When students engage in "retrieval practice," that is, taking practice tests, it benefits their memory retrieval even under stressful situations (Smith, Floerke & Thomas, 2016). While there is a DET practice test, written and spoken portions are not recorded. If DET examinees could access their DET recordings after each practice test, they could

not only experience taking the DET but see their efforts in context which could then guide their future practice.

Likewise, if examinees or their language instructors, could view their speaking and writing portions afterwards, they could undergo reflective techniques or receive detailed feedback about how to improve (Richards & Lockhart, 1996). Initially, another component of this study intended to provide participants this opportunity to view their speaking and writing portions with the author providing detailed feedback and guidance. Drawing from Jan Chappuis' advice, students would self-assess their DET performance prior to meeting with the author, then view and discuss the speaking and writing portions of the test together (Chappuis, 2015). Unfortunately, although this element was offered to all participants, at the time of this publication none of them have sought such feedback. This could be due to the limited amount of time left in the academic year, or individual's comfort level with seeing their English performance alongside the author. Nonetheless, educators must keep in mind the capacity for utilizing the DET's technology as a way for expanding students' testing knowledge and experience.

Limitations

Examining a test such as the DET with participants early in their university career held many challenges. One of the most obvious limitations to this study was the small sample size. There were several reasons for this including funding and the overall size of the Global Communication Department. Funding for this project was provided by the author's research budget and a grant provided by Hiroshima Bunkyo Women's University.[1] Since the price for each DET is $49/ 5900 yen, the sample had to be limited according to the available funds. Secondly, the small size of the current Global Communication department first and second year cohorts directly affected the candidate pool. At the time of Phase 1 in January 2017 the entire first-year cohort totaled only 18 students, of which 9 took part in the study. Likewise, the current first-year cohort is 31 of which fifteen participated in the study. Therefore, though several factors contribute to this small sample size, the study managed to receive its members from half of each cohort.

Another major limitation was the identification requirements in place by Duolingo. Participants cannot take the DET without a valid form of government ID which must contain a photo (e.g. passport or driver's license). This affected the study's candidate pool, particularly in Phase 1, because many of the interested would-be participants did not have the necessary identification to take the test. Therefore, one of the major reasons for conducting this study at the end of each academic year was to give participants time to obtain a passport or driver's license.

Heightened test security by Duolingo also affected this study. During the second phase of this project six students had their initial tests rejected due to excessive noise or not producing enough language during the speaking prompts. Another three students needed to resubmit their

1   The author has no affiliation with Duolingo and received no compensation in this study.

government ID before their scores could be released. Although these security actions are understandable in order to maintain the high standards of the DET, its effects did not go unnoticed by the participants. Of the six rejected participants, four decided to take it again, with one of them being rejected a second time. Since all the participants in the study lacked access to proper technology, accommodations had to be made during the limited period of time the study's computers were available. This resulted in noticeable frustration and anxiety from the affected participants. However, perhaps this negativity was only temporary because in the survey only one student indicated an unwillingness to try the DET in the future and another recommended securing a quieter location for future DET testing.

The change in and lack of access to the newly required technology also inhibited this study. In the first phase, students could easily locate, download and familiarize themselves with the DET app since each student at the university has an iPad mini. However, with the cancelation of the DET app participants had to use a computer with a web camera/microphone and connect using the google chrome internet browser. In addition, using a computer keyboard instead of the touch keyboard common on mobile devices seemed to have a negative effect on some. Specifically, during the interview several reported being apprehensive at their slow typing skills and recommended for future students to familiarize themselves with a computer keyboard prior to the DET.

A final limitation to this study involved a lack of private spaces needed to successfully complete the DET. The majority of the tests were administered in small, enclosed booths within the university's SALC. These rooms were not sound-proof and located directly next to an area dedicated for casual English chatting. Phase 1's timing in late January, paired with using the microphone within the iPad mini allowed participants to easily hear and complete their DET without complications. However, the second phase of the study had a nosier SALC, combined with the more sensitive microphones within the computers resulting in more tests being rejected. Private, quiet spaces on any university campus is difficult to find as was the case during the second phase of this project.

Whether this project will continue still must be determined. All participants involved in the second phase of this study lacked the technological or environmental requirements needed to complete the current DET. The hardware and software had to be procured by the author and quiet, solitary spaces needed to be located. Nonetheless, undergoing the DET does seem to have a positive effect on participants' motivation for continuing English studies and testing. Despite the limitations of this study, participants indicate a positive view towards the DET and its potential for future use as an English language proficiency tool.

Bibliography

Brenzel, J. & Settles, B. (2017). *The Duolingo English Test – Design, Validity, and Value*. DET Whitepaper (short). Retrieved from https://s3.amazonaws.com/duolingo-papers/other/DET_ShortPaper.pdf

Brenzel, J., Settles, B., & Tousignant, J. *A Look Inside the Duolingo English Test and New Features* (Webinar Video). Retrieved from https://englishtest.duolingo.com/resources

Chappuis, J. (2015). *Seven Strategies of Assessment for Learning* (2nd ed.). Boston, MA: Pearson Education, Inc.

Ishikawa, L., Hall, K., & Settles, B. (2016). *The Duolingo English Test and Academic English*. Duolingo Research Report DRR-16-01. Retrieved from https://s3.amazonaws.com/duolingo-papers/reports/DRR-16-01.pdf

Richards, J. & Lockhart, C. (1996). *Reflective Teaching in Second Language Classrooms*. New York, N.Y: Cambridge University Press.

Smith, A., Floerke, V., & Thomas, A. (2016, November). Retrieval Practice Protects Memory from Acute Stress. *Science* 25, Vol. 354, Issue 6315, 1046–1048. Retrieved from http://now.tufts.edu/news-releases/practice-testing-protects-memory-against-stress

Ye, F. (2014). *Validity, Reliability, and Concordance of the Duolingo English Test*. Technical Report, University of Pittsburgh. Retrieved from https://s3.amazonaws.com/duolingo-papers/other/ye.testcenter14.pdf

APPENDIX A

Interview Questions

A discussion immediately following the DET. Students could confer with others.

• What are your impressions of the DET?

• What did you like? Why?

• What did you dislike? Why?

• What was easy for you? Why?

• What was difficult for you? Why?

• What do you think you score will be?

• How will you use this information in the future?

• Would you like to try this test again in the future?

• What is your advice about the DET?

• Is there anything else you would like to discuss?

APPENDIX B

Phase 2 Survey

Anonymous and completed AFTER receiving DET scores. All questions were provided in English and Japanese.

1	Research consent
2	Which year are you? I am a first year student / I am a second year student.

3.	Please indicate your familiarity with the Duolingo English Test prior to this survey.
	    Very familiar (I have used it before)
	    Not so familiar (I have heard about it but never used it)
	    Not at all familiar (I have never heard of it or used it)

4.	Please indicate your satisfaction with the Duolingo English Test.
	    Very satisfied
	    Satisfied
	    Somewhat satisfied
	    Somewhat dissatisfied
	    Dissatisfied
	    Very dissatisfied

5.	Do you agree with your Duolingo English Test score?
	    Strongly agree
	    Agree
	    Disagree
	    Strongly disagree

6.	What do you like about the Duolingo English Test?
7.	What did you find difficult about the Duolingo English Test?
8.	How will the Duolingo English Test help you in your English language studies in the future?
9.	How likely are you to continue using the Duolingo English Test in the future?
10.	Do you have any suggestions regarding the Duolingo English Test at the BECC?
11.	Please use the space below to add any additional comments you would like to share.

APPENDIX C

DET Score Comparison

Based on the research that correlates Duolingo English Test scores to TOEFL iBT® and IELTS® Academic, the score converter below can help compare results.

DET Score	CEFR	TOFEL	IELTS
0–25%	A1/A2	n/a	n/a
26–29%	A2	60–62	4.5
30–38%	B1	63–70	5
39–49%	B1	71–81	5.5
50–60%	B1/B2	82–91	6
61–68%	B2	92–97	6.5
69–78%	B2/C1	98–104	7
79–86%	C1	105–109	7.5
87–93%	C1	110–115	8
94–97%	C2	94–118	8.5
98–100%	C2	119–120	9

Author's Note: The scores depicted above have been condensed due to space. For full score comparisons see https://englishtest.duolingo.com/scores

APPENDIX D

DET Score Interpretation

The Duolingo English Test is designed to assess real-world language ability, by measuring reading, writing, speaking, and listening skills. Each test taker's proficiency is reported as a holistic score on a scale from 0−100. The scores align to the Common European Framework of Reference (CEFR), an international standard of language proficiency.

The test scores may be interpreted as follows:

Score	Level	Abilities	Example
0%−16%	beginner	Can only understand very basic words or phrases in the language.	Can read public road signs, ask for basic directions, and fill out a simple form.
17%−30%	elementary	Can deal with simple, straightforward information and express themselves in familiar contexts.	Can have a short, coherent dialogue on topics of interest, but not extended conversations.
31%−50%	intermediate	Can understand the main points of concrete speech or writing on routine matters such as work and school. Can handle most situations that would come up while traveling where the language is spoken. Can describe experiences, ambitions, opinions, and plans, although with some awkwardness or hesitation.	Can open a bank account, if the procedure is fairly straightforward.
51%−70%	advanced	Can fulfill most communication goals, even on unfamiliar topics. Can understand the main ideas of both concrete and abstract writing, and interact with native speakers fairly painlessly.	Can show visitors around and lead a detailed guided tour of a place.
71%−90 %	proficient	Can understand a variety of demanding texts and conversations, also grasping implicit or figurative meaning that is hidden. Can use language flexibly and effectively for most social, academic, and professional purposes.	Can get and hold onto his or her turn to speak at a party, or respond to interrogating questions with little or no hesitation.
91%−100%	expert	Can understand virtually anything heard or read, even intellectually demanding material such as an academic lecture or a book on philosophy. Can use the language fluently and spontaneously in a way that can even be more advanced than an average native speaker.	Can scan long texts for relevant information, and differentiate finer shades of meaning in complex social and professional situations.