Gauging Extensive Reading's Relationship with TOEIC Reading Score Growth

Nathaniel Carney
Kobe College

This paper reports on a study relating the extensive reading achievement of an intact group of EFL learners at a Japanese university to the change in their institutional TOEIC reading scores after a period of seven and a half months. Similarly to other studies using inferential statistics to determine how extensive reading affects or relates to TOEIC scores, this study found almost no statistically significant relationship between increased reading and improvement in TOEIC reading scores. Likewise, the extensive reading group did not have significantly higher TOEIC reading scores than other similar proficiency groups at the same university who were not doing extensive reading. In response to this and other studies' results, the paper includes an extended discussion regarding the plausibility of researching extensive reading's relationship with TOEIC scores and important considerations for such research if it is carried out. The paper concludes with a call for more widespread collaboration among extensive reading researchers.

Carney, N. (2016). Gauging Extensive Reading's Relationship with TOEIC Reading Score Growth *Journal of Extensive Reading 4*(4). http://jalt-publications.org/jer/

Keywords: extensive reading; TOEIC; reading; standardized tests

Introduction

Extensive reading (Susser & Robb, 1990) is enthusiastically promoted by a number of L2 reading researchers (Grabe, 2009; Krashen, 2004; Mason & Krashen, 1997; Nation, 2008; Nuttall, 2005) but despite having grown in popularity, there still seems to be resistance to fully incorporating it into English teaching in secondary and higher education curricula (Grabe, 2009). Various reasons English L2 programs might have for not doing extensive reading have been discussed (Grabe, 2011; Macalister, 2010; Renandya, 2007), and one of these reasons concerns the lack of convincing research showing how extensive reading leads to measurable gains in English proficiency. Recent research under controlled conditions has yielded some evidence that extensive reading can benefit general reading ability (Yamashita, 2008) and reading fluency (Beglar, Hunt, & Kite, 2012). If such benefits are possible, given a certain amount of extensive reading done over a given time, one wonders if the acquired proficiency will show meaningful effects on other measures of proficiency, specifically high stakes standardized reading tests.

Those selling books about extensive reading (e.g., Furukawa, 2010) or textbooks for it (McDonald & Malarcher, 2008) clearly imply that extensive reading will improve test scores. Nevertheless, there is no current research showing that this is true. In this paper, I investigate this concern through examining the relationship between extensive reading and reading scores on a standardized English proficiency test, the institutional TOEIC (Test of English for International Communication). After presenting the results, I engage in an extended discussion about the relationship between extensive reading and TOEIC scores.

The TOEIC in Japan

The TOEIC is one of the most common standardized English proficiency tests in the world. Japan and South Korea are the two countries where the TOEIC is taken the most (Chapman & Newfields, 2008). The TOEIC's widespread use in Japan is not coincidental; it was developed specifically for Japan, by the Education Testing Service (ETS) in the late 1970s.

TOEIC, along with the easier TOEIC Bridge version, serves numerous purposes in higher education and work environments in Japan (ETS, 2013a; Tokunaga, 2008). For example, at some Japanese universities students are streamed by TOEIC scores (Takahashi, 2012), they qualify for certain programs (e.g., honors programs) via TOEIC scores, and they may even win scholarships or be accepted for study abroad because of TOEIC scores. Also, students wishing to become certified teachers may be required to attain certain TOEIC score levels (Chapman & Newfields, 2008). After leaving university, TOEIC scores again may be used at Japanese companies for hiring decisions, in assigning overseas work posts, and even to help determine salary or bonuses. Softbank, for instance, reportedly has a system of paying employees a one million yen bonus if they score at least 900 points on the TOEIC (Ryall, 2013). Furthermore, the TOEIC shows no signs of disappearing, with a record number of companies using the test in 2011 (ETS, 2011). Given the popularity and importance of the TOEIC, some universities even offer TOEIC classes to help learners improve their scores. Part of the TOEIC test is a reading section. Since extensive reading is proposed to benefit reading fluency (e.g., Grabe, 2010), and better reading fluency would seem to improve performance on a reading comprehension test, it is natural to investigate the relationship between the two[1].

Extensive Reading and TOEIC relationships

There have already been efforts to examine the relationship between extensive reading and TOEIC score improvement in Japanese university contexts. Storey, Gibson, and Williamson (2006) found no statistically significant improvement of reading scores between an experimental and control group of second and third year L1 Japanese science students. However, through closer analysis, the authors did find that in the experimental group, those that read more had a 30% greater improvement in TOEIC scores relative to those who read less. Nishizawa, Yoshioka, and Fukada (2010) and Mason (2011) reported stronger implications that extensive reading led to TOEIC score increases. Nishizawa et al. (2010) conducted a longitudinal study with *kosen* (a school of technology for students 15-20 years of age) engineering students over a four-year period. The authors found that students who read a great deal seemed to increase TOEIC scores significantly. Specifically, they noted that after reading around 300,000 words students' scores begin to increase noticeably. While inferential statistical measures were not used to detail such increases, the general data presented suggested the potential relationship the authors proposed. Mason (2011) did a case study of an L1 Japanese adult hospital employee, age 42, who did extensive reading for a year under the researcher's guidance. Mason remarked that the man improved his TOEIC score by 180 points after doing a large quantity of graded reading (i.e., 2590 pages), though few details about the contribution of other possible moderators were reported in the study. Finally, O'Neill (2012), with a sample size bigger than any of the studies reported above (*n* = 213), found no statistically significant improvement between a group of students who did an extensive reading program and a group that did not.

Taken together, these studies offer mixed results, and as with much L2 reading research, the studies employed different methods for gathering data, analyzing data, and also were analyzing different groups of learners. Two studies employing inferential statistics (O'Neill, 2012; Storey et al., 2006) found no statistically significant gains. At the same time, in three studies (Mason, 2011; Nishizawa et al., 2010; Storey et al., 2006) authors suggested strongly that larger amounts of reading did seem to improve TOEIC scores, while in another study (O'Neill, 2012) the author was not able to test whether that might be the case because of logistic difficulties. In short, researchers have proposed a likely connection, but definitive evidence remains elusive.

Methodology

Participants

Participants in this study were 20 female first-year English majors at an all-women's 4-year liberal arts college in Japan. All students were native speakers of Japanese, from 18 to 19 years old over the course of the study. Seven of the students had short-term (one month or less) study abroad experience in native-English speaking countries. All students were relatively similar in terms of initial TOEIC scores, within a range of 70 points. While university Japanese English language learners are sometimes regarded as having low motivation, the group of learners in this study were generally highly motivated in the instructor's opinion. This impression was supported by their high attendance rate (94% of classes attended) and active participation in class. This was not unexpected given that they were one of the higher-level classes at the university, and English was their chosen major.

Procedures

This study was conducted in a university context where extensive reading has not systematically been a part of the university English curriculum. In this case, "not systematically" means that, except for the one class focused on in this paper, none of the other classes had an extensive reading program in which students were required to do large amounts of graded reading over a defined period of time. A variety of student performance data were collected as a normal process of the class. This data included the number of words read through extensive reading, recorded through the online M-Reader website, and usage statistics from Word Engine, an online vocabulary learning system.

Extensive reading for the class comprised

20% of the final class grade. Generally, extensive reading was additive (Robb & Kano, 2013), that is, it was done outside of class time as homework. However, students did read in class for 15 minutes during both the second and third class meetings. Extensive reading was assessed as the number of words read at an appropriate level for the student, and these are the data used to measure extensive reading in this study. Word counts of extensive reading were calculated via the M-Reader website. On the website, when students pass a quiz on a book (i.e., score 60% or higher), the exact or estimated word count for the book is added to their reading total. It is recognized that the word counts in M-Reader will not be exact given that some students took quizzes but did not pass them, hence they did not receive credit for reading those books even though they might have. Similarly, there could have been cases where students were able to pass quizzes without fully reading books. Despite these imperfections, the word counts were reasonably accurate estimations. Another imperfection of using an online assessment system like M-Reader is the possibility of cheating. However, the likelihood of widespread cheating was very low. Student progress was monitored weekly, and students had a relatively high interest in English study. As well, M-Reader was only being used in this one class at the university, so the number of partners with whom students could possibly cheat was quite limited. Finally, M-Reader has two built-in cheating monitoring systems that allowed the researcher to verify that at least certain types of cheating (e.g., taking tests at almost the same time, reading many of the same books as another student) were not occurring.

Initial extensive reading levels were determined during the first class meeting through students' self-assessment of their understanding of a few pages of reading from a level 2, a level 3, a level 4, and a level 5 Oxford graded reader. Students were directed to choose a level in which there were no more than five unknown non-proper nouns on the page. If they found the level to be too easy or difficult after reading part of the book, then it was suggested they choose the next higher or lower level respectively. Students chose their book levels throughout the study period, and all books read by students were from the college library, which had a collection of over 400 graded readers from a variety of publishers. Students' choice of books was easily tracked through the M-Reader online system, and no students were found to be reading books at levels far from what the instructor considered appropriate. Students generally read between one to three books per week over the course of the seven-and-a-half month study period, taking a quiz on the M-Reader site for each book read. There were five cases (out of more than 300 books read) in which books had no quizzes on M-Reader. In these cases, students submitted 150 to 200-word book reports online, in English, summarizing the story and giving a short reaction. Thereafter, they brought the book to class where the instructor quickly looked through the book and asked the student to answer two or three general questions about the book to verify they read and understood basic details. For reports, the instructor manually input the word counts for the books into students' M-Reader accounts.

Word Engine use in the class comprised 10% of students' final grades. Except for the class time used to set up and orient students to the program, Word Engine study was all conducted outside of class. Word Engine is a spaced-repetition online, mobile-enabled vocabulary flash card program. Word Engine data for this study was defined as the number of words learned, information obtainable through the Word Engine management system. Word Engine study can be done on computer, or, as became available during the course of this study, on mobile devices like smartphones. On the Word Engine website (www.wordengine.com), users take an initial vocabulary size test to determine their level. Users must then create an account on the system which allows them to study vocabulary. The system automatically tailors the vocabulary words and phrases for each user according to their vocabulary level, determined by the initial test. As detailed by Cihi, Browne, & Culligan (2013), in Word Engine users are presented with 68% known words when they initially begin using the system; this allows learners to gain confidence as well as to fortify their knowledge of important, frequent lexical items. Gradually the system introduces a higher ratio of words the learners would be unexpected to know based on their initial vocabulary test score. The format for study in Word Engine is short quizzes, some in game-like formats, and includes audio of the words being studied. After answering correctly about a certain vocabulary item a number of times, it is recorded as "learned" by the system. These "words learned" constitute the data used to measure vocabulary learning in this study. Limitations of using "words learned" data should be noted. For one, the initial vocabulary size test on which Word Engine bases the words given to users is not necessarily reflective of student levels. So, if a student does poorly on the test despite having a large vocabulary, Word Engine will present them overwhelmingly with many words they already know, and then it will be easier for them to "learn" those words. Despite this drawback, there was no evidence that students in the class tried to do poorly on the quiz. This was further unlikely since no other students at the university were using the system, and the students did not know how the system worked until after taking the quiz.

Word Engine data for this study came from participant use over the course of 23 weeks, from mid-May until the end of October, with an optional two-week break during the summer (so the required number of weeks was 21). Students were assigned a weekly goal of answering 100 vocabulary questions correctly on the Word Engine system. While most of the participants in this study did Word Engine consistently almost every week during the 23-week period in which it was assigned (*M* = 17.45, *SD* = 4.88), there was a wide range of practice. One student did Word Engine for only three weeks, while three students did it for twenty-two weeks. Since those who missed weeks using Word Engine generally had lower "words learned", the use of "words learned" data generally seemed to represent student gain from the system well, but it is recognized that consistent study could result in qualitatively different learning than inconsistent study. While this issue is brought up again later in the discussion, the point now is that the Word Engine data, like the extensive reading data, are imperfect.

The research questions for this study were as follows:

1. Do increasing levels of extensive reading over a relatively short time (seven and a half months) correlate with higher TOEIC reading scores?

2. If so, to what degree are such relationships significant compared with other moderators, such as vocabulary study?

3. Are the changes in TOEIC reading scores of the extensive reading group significantly different from similar proficiency groups that did not do extensive reading?

In order to answer the research questions, extensive reading data and Word Engine data were compared with changes in learners' IP-TOEIC reading scores from April and November. These two factors, extensive reading and Word Engine, were chosen as they represented two variables that would likely be related to reading success: reading skill and fluency on the one hand, and vocabulary on the other (Grabe, 2009). After determining relationships between TOEIC change, Word Engine and extensive reading, changes in TOEIC reading scores from the 20 learners were compared with TOEIC reading score changes of other similar level learners (i.e., other intact classes) in the same year who did not do extensive reading, and also with similar level learners from the previous year who did not do extensive reading. Results from these analyses are presented in the following section.

Results

Means and standard deviations for the amount of extensive reading done by participants and their usage of Word Engine are presented in Table 1. While TOEIC descriptive data are not included in Table 1[2], reading score changes ranged from slightly negative to almost 100 points gained.

All statistical results were computed using SPSS, version 20. First the data were screened for missing values or errors, and then scatterplots of the data along with skewness and kurtosis statistics were examined to determine whether the data were normally distributed. The data were basically normally distributed, but, following Field's (2013) suggestion, bootstrapping was used as a robust method of dealing with small samples and calculating bias corrected and accelerated bootstrap 95% confidence intervals. Thus, in order to answer the first and second research questions, bootstrapped Pearson r correlations along with bias corrected and accelerated bootstrap 95% confidence intervals were computed for extensive reading, Word Engine, and participants' change in TOEIC score from April to November. Results are shown in Table 2.

The first research question in this study inquired whether there was a significant correlation between extensive reading and the change in TOEIC reading scores over the course of the extensive reading program. The second question concerned how that relationship compared with that of other moderators, in this case, Word Engine vocabulary study. As shown in Table 2, there was a statistically significant correlation between extensive reading and change in TOEIC reading scores, but the correlation was low. There was similarly a slightly significant correlation between

Table 1
Descriptive Statistics for Extensive Reading and Word Engine

	N	Mean	SD
Extensive Reading: words read	20	195451	90925
Word Engine: words learned	20	448	133

Table 2
Relationships Between Extensive Reading, Word Engine, and TOEIC Reading Score Change

Measure	1	2	3
1. Extensive Reading	___		
2. Word Engine	.50* [.02, .78]	___	
3. TOEIC reading score change	.48* [.01, .77]	.36 [-.01, .65]	___

* p < .05

extensive reading and Word Engine. The broad bootstrapped confidence intervals of the relationships suggest that the relationship is neither strong nor clear. If a conservative approach is taken and the Bonferroni correction is applied because multiple correlations are being computed (Field, 2013), then α = .016 and neither of the statistically significant correlations above would be significant. In short, it is not possible to completely discount the statistically significant finding above, but a stronger relationship between TOEIC reading score change and extensive reading would be more convincing. Word Engine's correlation with extensive reading deserves some interpretation as well. A rationale for including the Word Engine and extensive reading relationship was to consider whether such a relationship could indirectly signify student motivation. In other words, as extensive reading and Word Engine were both assessed elements of the course, and the best students often do their homework and also study hard for the TOEIC, finding a relationship between extensive reading and Word Engine would be expected. On the other hand, Word Engine is clearly not correlated with the change in TOEIC reading score change. Does this signify that extensive reading indeed may be more strongly related to TOEIC score change than Word Engine? As will be addressed in the discussion, small sample sizes and relatively short-term studies like this unfortunately can leave more questions than answers. While the relationships here can be interpreted in varied ways, what might be most salient is their general weakness, suggesting either a lack of meaningful relationship, or the need for broader and more longitudinal studies.

The third research question asks whether the extensive reading group's TOEIC reading score change differed from other groups of similar English proficiency. To answer this question, independent t tests were used to determine whether there were any significant differences. TOEIC reading score change in the intact extensive reading group was compared with that of two other similar level intact groups in the same academic year that did not do extensive reading, but did other activities (e.g., textbook reading skills activities, reading books as class readers, intensive reading of difficult texts). As well, TOEIC reading score increase comparisons were made between the extensive reading intact group and a comparable proficiency level non-extensive reading group (not an intact group) from the previous year and an intact group from the previous year that was nominally at the same level (i.e., had the same level name) as the extensive reading group. The reason these two previous year groups were chosen was because the system for forming classes the previous academic year was slightly different (i.e., they were not simply streamed by TOEIC). In short, the overall idea was to compare groups of similar proficiency that had not done extensive reading with the group that had. Before conducting tests, it was confirmed that the data sets met required *t* test assumptions of normality and homoscedasticity. The results are found in Table 3. The *n* sizes for each of the comparison groups were as follows: higher level *n* = 22, lower level *n* = 23, same class *n* = 25, same level = *n* = 26.

Table 3 compares the change in TOEIC reading scores from non-extensive reading groups with the extensive reading group focused on in this study. The “same year” groups are two intact classes that had a slightly higher English proficiency level (i.e., “higher level”) and a slightly lower proficiency level (i.e., “lower level”).

Because classes were streamed by proficiency level (i.e., TOEIC scores), these were the two intact classes closest in proficiency to the class that did extensive reading. From the previous year, two groups were selected for comparison. As explained above, because the previous year had not been streamed solely by TOEIC scores, one group chosen for comparison was nominally the “same class” as the extensive reading group. In other words, the class prefix used to denote their level was the same, so they would be roughly the same proficiency level. The other group noted as “same level” was a group of students, whom, if they had been streamed by TOEIC scores, would have been at the same average proficiency level as the extensive reading group. So, the “same level” group of students were not actually in the same intact class, but rather they, as a group, had the same TOEIC proficiency levels as the extensive reading group. As mentioned, the point of the comparison is to compare change in TOEIC among groups of similar proficiencies, so that the TOEIC reading change scores of the extensive reading group in this study can be contextualized as especially meaningful or not.

As Table 3 shows, in no case was the TOEIC reading score change of the extensive reading group statistically significantly different than that of the comparison groups. Cohen’s *d*, a standardized effect size value, shows that the largest differ-

Table 3
Comparison of ER-Group and Non-ER Groups’ TOEIC Reading Change

Comparison non-extensive reading groups	*t*	*p*	Cohen’s *d*
Same year non-ER group (higher level)	t(40) = .94	.36	.30
Same year non-ER group (lower level)	t(41) = -.26	.80	-.08
Previous year non-ER group (same class)	t(43) = -1.90	.06	-.58
Previous year non-ER groups (same level)	t(44) = -1.53	.13	-.46

ence was between the previous year's 'same class' (nominally the same class) and the extensive reading group. The fact that the Cohen's *d* value is negative means that the extensive reading group's TOEIC reading change score was higher than that of the previous year's 'same class', which is the direction an extensive reading promoter would expect to see in the relationship. While this is mostly the case, it can be seen that the same year's 'higher level' class actually had a positive Cohen's d value, meaning they had higher TOEIC reading change scores than the extensive reading group. In any case, it is essential to note that none of the differences are statistically significant, so the important conclusion of the *t* test results is that there was no difference in TOEIC change scores between students who did do extensive reading and those who did not.

Discussion

In this study, only a marginal statistically significant relationship was found between extensive reading and TOEIC reading score increases. In Storey et al.'s (2006) and O'Neill's (2012) studies, no statistically significant relationship was found between the amount of extensive reading done and TOEIC reading score increases. It is hoped and believed that extensive reading over time would benefit standardized test reading scores, like the TOEIC. However, despite the important limitations notable in this study (e.g., the small sample size and the inability to include consideration of all possible moderators affecting TOEIC score changes), this study adds to the research showing a murky picture of how extensive reading might relate to TOEIC scores. In other words, findings reflect those of other studies in which the relationship between TOEIC scores and extensive reading are largely inconclusive in regard to any benefit. To be sure, there might be many researchers who would completely expect such unclear findings. Thus, the discussion that follows is meant to address the findings of this study, as well as the body of research which to this point has been relatively unclear about whether extensive reading has a measurable effect on TOEIC score growth. This discussion will address some important questions that researchers should consider when examining links between extensive reading, the TOEIC, and standardized tests in general. While the discussion that follows does relate to the study reported on in this paper, another important goal of the discussion is to point a way forward for future research. To frame the discussion, the following four questions will be discussed.

Question 1: *Are TOEIC reading scores a poor measure of reading and therefore not a worthwhile focus of extensive reading research?*

Question 2: *If extensive reading does positively affect TOEIC reading scores, is it possible to have a research study controlled enough to show it?*

Question 3: *If extensive reading is done often enough and for a longer period of time, will TOEIC scores increase?*

Question 4: *Will a well-designed extensive reading program increase TOEIC reading scores even when a poorly designed program does not?*

Each of the questions above will be discussed in turn, relating the answers to the study presented here as well as to other previous research.

Question 1: *Are TOEIC reading scores a poor measure of reading and therefore not a worthwhile focus of extensive reading research?*

The study in this paper searched for a relationship between extensive reading (operationalized as the number of words read at an appropriate level) and TOEIC reading score growth. However, if one answers 'yes' to Question 1 above, then the study presented in this paper, and any research looking for extensive reading's effect on or relationship with TOEIC would be without purpose. Answering 'yes' to Question 1 implies a general low regard for the TOEIC reading test as a measure of reading ability, and it is supported by some research. Chapman (2003, 2006) and Childs (1995) have criticized the use and structure of the original TOEIC of listening and reading, and Chapman and Newfields (2008) discuss various issues with the new listening and reading TOEIC. Specifically regarding reading, Chapman and Newfields (2008) suggest that gap-fill and multiple choice methods of testing reading do not match the authentic ways in which people read texts. The authors do not elaborate on the meaning of this point, but it might be inferred that, since the authors cite Alderson (2000) in that section of their paper, they might concur with Alderson's assertion that "the challenge for the person constructing reading tests is how to vary the reader's purpose by creating test methods that might be more realistic than cloze tests and multiple-choice techniques" (Alderson, 2000, p. 249). Whether this is a fatal flaw of the TOEIC reading test is undetermined. The TOEIC reading test does succeed in presenting certain realistic textual forms (i.e., business letters, memos, emails, and websites) and both single and double passages in the multiple choice sections, though this of course only comprises 48 of the 100 points on the reading test. The remainder of the test consists of 12 points in the cloze text completion section and 40 in the incomplete sentences multiple choice section. In terms of validation, ETS has supported some research of the reading section (ETS, 2013b) to backup claims of validity.

Grabe (2011) raises a different problematic issue about using standardized tests to measure extensive reading effects. He cautions that the implicit learning that should result from extensive reading is not measured in reading tests that do not use time-constrained tasks (Grabe, 2011). W. Grabe (personal communication, February 11, 2014) explains this concern as follows:

> A true time-constrained task assumes that a larger percentage of students may not fully finish the task, or that the task measures the time taken toward completion. The point is that the task is intentionally designed to create a high level of time pressure and that is part of the scoring assumption.

So, while the TOEIC reading section is timed, that time is not accurately figured into the test taker's score. That would be impossible given the structure of the TOEIC, where guessing is not penalized and there is no way to know from item to item how much time is spent on each. A true time constrained task would give higher points to a fast test taker answering correctly versus a slow test taker who answers correctly.

These concerns about the value of the TOEIC reading test and standardized testing's relationship to extensive reading are important and justified. However, just as extensive reading intuitively makes sense as a method for gaining reading proficiency in a language, many would intuitively argue that extensive reading should improve performance on almost any test involving reading in that language. With extensive reading purported to improve fluency and the automatic processing of language (Grabe, 2010), it makes sense that learners who have done enough extensive reading would be more comfortable and faster at processing the language in the reading section of the TOEIC, even given that TOEIC reading and extensive reading potentially have very different purposes and lexical densities. With faster processing, one would assume that eventually extensive reading will lead to greater TOEIC success. If one agrees with this point, then the second, third, and fourth questions posed above become relevant.

Question 2: ***If extensive reading does positively affect TOEIC reading scores, is it possible to have a research study controlled enough to show it?***

The second question raises a challenging issue of researching extensive reading. There are many critiques of L2 extensive reading research even from those who promote it (Beglar et al., 2012; Grabe, 2011; Robb & Kano, 2013; Robb & Susser, 1989; Susser & Robb, 1990; Yamashita, 2008). One issue in many research studies is the lack of accurate or appropriate instruments for measuring language change or gathering data. For example, in the author's study presented in this paper, M-Reader has faults as the sole indicator of how much participants were reading, and Word Engine has problems as a sole measurement of how much vocabulary growth has occurred. Another typical critique in extensive reading research is the lack of controls on variables besides extensive reading which could influence outcomes. The study presented in this paper did suffer from a certain lack of controls. For example, though unlikely, it could not be guaranteed that students in the non-extensive reading comparison groups for this study were not doing a great amount of reading beyond their coursework. In traditional experimental research, all variables that might affect outcomes in an experiment should be controlled or accounted for, and the sampling of participants should be randomized. Of course, in most second language acquisition research, the reality is usually more of a compromise (Mackey & Gass, 2005), and this can certainly be said of extensive reading research. Given this compromise, there is always criticism that can be leveled at studies that employ inferential statistics to determine outcomes. This might suggest to some that qualitative studies are a better route for extensive reading research. Indeed, this is a revealing avenue (Judge, 2011; Pigada & Schmitt, 2006). However, it also might be said that well-designed quantitative studies with imperfect controls and intact classes are useful if they can give general impressions of what might be happening (Mackey & Gass, 2005). An excellent recent example is Beglar et al. (2012), who managed to create a well-designed study within the context of a university curriculum. The study, looking at the effects of extensive reading on reading fluency, had a reasonably large number of participants (*n* = 97) and tracked their progress over a reasonably long period of time (28 weeks). More significantly, the study involved a control group, gathered specific information about the level and quantities of all readings done by participants in the study, and had pre and post-tests measuring reading speed which had been piloted and analyzed for item reliability before their use in the study. In short, the study was thorough in using reliable measuring instruments and clearly describing materials and procedures in the study. Another study design to consider is Robb and Kano (2013), who examined the performance of a large group of participants' (i.e., more than 2000 students) performance on a university-designed reading test. The participants were of two different university cohorts, one that had done extensive reading for an academic year and one that had not, and their reading test scores were compared. Using similar logic to that used in this paper, the non-extensive reading comparison group was not a control group per se, but it was simply a group that had not done the extensive reading program. The extensive reading group had convincingly stronger, and statistically significant, performance on the reading section of the university-designed exam. If such a study had been conducted using TOEIC reading tests as the pre/post-test instead of a university-designed test, then it would ideally show what studies thus far have not been able to convincingly do. As Robb and Kano commented, they did not have to control for many variables in their study since they were simply making comparisons of two groups that, aside from extensive reading, would experience virtually the same English curriculum and exposure. In the absence of such a fortunate coincidental design, however, designing a study that can reveal whether TOEIC scores are being affected by extensive reading can be quite challenging. In a typical L2 English curriculum in Japan, for example, learners will be taking a number of classes in English, and possibly even more than one reading class. They will have different instructors, and they will be in classes alongside learners of varied proficiencies. Thus, even if the TOEIC scores of a group doing extensive reading increase significantly more than those of a group not doing extensive reading, how can a researcher possibly claim that the cause was extensive reading? Some would say that it is just not possible to design a study reliable enough. Others would say that if enough data were collected on all other moderating variables, or if a new extensive reading program were introduced to an otherwise unchanged curriculum (e.g., like Robb & Kano) then it is possible. Another perspective and possible fruitful path forward might be increased collaboration on studies, the advantages of which will be discussed further in the conclusion. Before that, the third question concerning program length is important to address.

Question 3: If extensive reading is done often enough for a long period of time, will TOEIC scores increase?

The third question above draws attention to how long extensive reading programs should be. The study in this paper lasted seven and a half months, but this was only because the instructor was able to assign reading through summer break and into the following semester. Like this, the length of an extensive reading program is often dependent upon how much control the teacher/administrator has over the program. Why does the length of programs matter? The beginning of this paper reviewed five studies that have looked for relationships between TOEIC and extensive reading. Three of these studies (Mason, 2011; Nishizawa et al., 2010; Storey et al., 2006) proposed that with large enough quantities of extensive reading, TOEIC scores increase. This sounds simplistic, but is an interesting conjecture. Nishizawa et al. have suggested that reading 300,000 words may be a sort of tipping point for TOEIC score increases. In the study presented here in this paper, only three students achieved this reading quantity. It is notable that all three increased their TOEIC scores far more than average, perhaps a statistically insignificant point given the small sample size, but of interest to those promoting extensive reading. The fact is, however, many extensive reading programs are not long enough or do not encourage frequent enough reading for 300,000 words to be read. The study reported on in this paper is such a case, where only three of twenty participants could reach that threshold. The point has been made before that extensive reading programs often end before they show results (Grabe, 2011; Grabe & Stoller, 2002).

Table 4
Projected amount of reading time required to reach 300,000 words

Length of Program (no. of semesters)	Reading speed (words per minute)	Minutes per day	Weeks (5 days per week)	Total Words
1	100	20	13	130000
1	150	20	13	195000
2	100	20	26	260000
2	150	20	26	390000
4	100	20	52	520000
4	150	20	52	780000

Nishizawa et al. reported that it took three years for 75% of learners to reach 300,000 words. If one were to make 300,000 words a goal, then how long would programs have to be? Taking the Japanese university context as an example, most semesters last fifteen weeks, with perhaps thirteen of those used for most of the coursework/assignments. Table 4 shows simple calculations about how much reading time would be required for learners to read a certain amount of words, only using actual class weeks (i.e., 13 per semester), based on slow (100 words per minute) and fast (150 words per minute) reading speeds.

Table 4 is meant to act as a heuristic for considering how long a program would need to be in order to read 300,000 or more words, but it is, of course an oversimplification of the real process. The author's students participating in this study spent time finding books, checking them out, and doing activities for evaluation (i.e., quizzes or book reports). Such work time is not included here. So, some teachers might feel that reading twenty minutes a day, five days a week during semesters is an ambitious expectation. In any case, Table 4 shows that a semester is basically too short for most students to reach 300,000 words in an extensive reading program. In a year-long program it might be possible, especially if the program follows a suggested guideline of having students do a significant amount of reading during class time (Takase, 2008). This discussion is all predicated, of course, on the proposed 300,000 word threshold for enjoying significant TOEIC benefits from extensive reading. Most recently, McLean (2014) has raised the issue of how much extensive reading should be done in a program. In terms of pedagogy, McLean proposes that extensive reading programs last for at least two years. I strongly support his pedagogical proposition. With a two-year extensive reading program, and reasonable expectations about reading speed (i.e., 100 words per minute), the majority of participants would be able to reach the 300,000 word threshold mentioned by Nishizawa et al. (2010).

Question 4: Will a well-designed extensive reading program increase TOEIC reading scores even when a poorly designed program does not?

Discussing research about extensive reading and the length of extensive reading programs coincides with the fourth and final question raised above; extensive reading program design may affect students' success – both in terms of student's enjoyment of the programs as well as the program's effectiveness. The study described in this paper was largely modeled upon that of other researchers, and student evaluations given at the end of the semester revealed that most students enjoyed the reading and autonomy offered by the extensive reading program. Enjoyment is quite important since extensive reading, or pleasure reading (Beglar et al., 2012), should be fun. Nevertheless, if programs are carried out over a long period of time (e.g., two years, as suggested above) and reading is a requirement, the issue of evaluation looms. All programs must consider how or whether to evaluate student progress, and evaluation can be seen as taking the fun away from reading. As efforts are made to measure and make students accountable for extensive reading through oral reports, written reports, discussion groups, online quizzes, and/or other activities, there is concern over whether the pleasure continues to exist. If students are required to read 300,000 words but don't enjoy it because of evaluation processes, will they still benefit? Concern over evaluation effects on extensive reading led Stoeckel, Reagan, and Hann (2012) to investigate whether a requirement to take quizzes for extensive reading books had a detrimental effect on students' reading attitudes. The authors had two extensive reading groups, one required to take quizzes (originally from an online extensive reading management system, The Extensive Reading Project at xreading.com), and one not required to take quizzes. The authors found no difference in reading attitudes between students required to take quizzes after reading and students who were not required to take quizzes. By no means does one study definitively answer the question (the authors point this out themselves), but it at least leaves the question open about whether broad, standardized evaluation of extensive reading is really bad. What is true is that any program interested in measuring student gains from extensive reading must have a well-designed, reliable system for gathering information about the quantity and quality of students' reading.

Another element of program design is the consistency with which extensive reading is done. In the study presented in this paper, the author had the advantage of teaching the same students in the first and second semester, so students continued reading during the two-month summer break. The M-Reader system used in the author's study allows administrators to limit how often students take quizzes on books, thus preventing students from trying to read all their books at one time. However, in the Japanese university context, and probably in many other contexts, there are long breaks between semesters in which it may be difficult for instructors to require reading. If students are not reading during these times, how does that affect potential benefits gained from reading? Likewise, when assigned to do reading, if students procrastinate and attempt to read many books during the final week before being evaluated, does it affect their reading benefits? These are questions that future research might focus on.

There are many other design issues to consider as well that could impact the benefit of an extensive reading program. Takase (2008) cites book levels and the amount of in-class reading as critical. The amount of instructor support introducing, discussing, and motivating students over the course of a program could be important. The ease with which students can get books and the number of books available could be important. Bringing the discussion back to how extensive reading may affect TOEIC scores, design issues suggest that it is not just the product of extensive reading (e.g., reading 300,000 words) that is important, but also the process i.e., a well-designed program.

Conclusion

The purpose of this paper has been to examine the relationship between extensive reading and TOEIC reading score growth. The author's own study suffers from familiar limitations of a small sample size and nongeneralizability, but has findings that concur with other recent similar research; there is no clear relationship between the amount of extensive reading done and TOEIC score growth. Nevertheless, given the widespread use of TOEIC and its potentially high stakes, effort has been made to discuss the plausibility of the notion that extensive reading could in fact noticeably benefit TOEIC scores under certain conditions. As with many papers, this one concludes with a call for more research in this area; in the author's opinion, extensive reading's relationship with TOEIC and other standardized test scores' change is an important area to be explored. However, rather than just a call to research, it is a call for research collaboration.

Collaboration on extensive reading research – where researchers at different institutions work together on a given project – is quite rare. With greater sample sizes representing different learning contexts, research relating extensive reading to TOEIC might become more convincing and ecologically valid, even with a given degree of imperfect controls on other moderators. Larger samples tend to better represent the overall population, thus offering results that are more generalizable. When the number of participants is large and also drawn from a variety of contexts, the evidence is strengthened even further. Collaboration also offers the benefit of shared expertise and effort among researchers and teachers. Some teachers/researchers may lack either the time or skill to properly analyze the effects of their programs, while others with the time and skill may be working in small extensive reading programs from which it is difficult to generalize. The time seems ripe for significant broad-based collaboration on certain critical issues related to extensive reading. Evidence concerning the beneficial impact of reading a lot (e.g., 300,000 words) would be extremely useful for program designers. It could provide significant evidence to programs where there is resistance to doing extensive reading for more than one semester or to including it as a mandated part of the school-wide curriculum. Collaboration among teachers, researchers, and administrators offers its own challenges, but it also offers great potential for better extensive reading research.

In conclusion, extensive reading seems to be a big topic in Japan and in L2 reading research in general. While Grabe (2011) asked, "Why isn't everyone doing it?", at least now, many programs in Japan are. So, the question I would like to ask is, "Why aren't we collaborating to research about it?" The question is not as catchy, but the purpose is clear. With leadership from extensive reading organizations (e.g., The Extensive Reading Foundation, www.erfoundation.org, and the JALT Extensive Reading Special Interest Group, http://jalt.org/er/) and more communication among researchers, better answers to big questions may become possible.

Notes

1. Upon discussing this study at a recent conference, one attendee suggested investigating the relationship with listening rather than reading. While I did not have a rationale for doing so, she suggested that there would be a strong relationship with listening, perhaps because of research on the importance of phonological awareness for L1 reading in young learners. At this point, I still regard the relationship of extensive reading to reading as being more salient and reasonable.

2. For institutional privacy purposes, descriptive data of TOEIC scores are not included in the tables in this paper. Contact the author for more information.

References

Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning, 62*, 665-703. doi: 10.1111/j.1467-9922.2011.00651.x

Chapman, M. (2003). TOEIC: Tried but undertested. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2) 2-7. Retrieved December 29, 2009 from http://jalt.org/test/cha_1.htm

Chapman, M. (2006, April). An overreliance on discrete item testing in the Japanese business context. Paper presented at the International Conference on English Instruction and Assessment. National Chung Cheng University, Taiwan. Retrieved from http://fllcccu.ccu.edu.tw/conference/2005conference_2/download/C07.pdf

Chapman, M. & Newfields, T. (2008). The 'new' TOEIC. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2), 32-37. Retrieved December 29, 2009 from http://jalt.org/test/cha_new.htm

Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In J.D. Brown and S.O. Yamashita (Eds.), *Language Testing in Japan* (pp. 66-75). Tokyo, Japan: JALT.

Cihi, G., Browne, C., Culligan, B. (March 2013). "V-Check and WordEngine Academic FAQ (Ver. 1.8)." *WordEngine.jp*. June 4, 2015. http://www.wordengine.jp/research/main

Educational Testing Service. (2011). *TOEIC® newsletter: No 118 – Digest Version*. Retrieved from http://www.toeic.or.jp/english/toeic/about/data/newsletter.html

Educational Testing Service. (2013a). *TOEIC® newsletter: No 119 – Digest Version*. Retrieved from http://www.toeic.or.jp/english/toeic/about/data/newsletter.html

Educational Testing Service. (2013b). *TOEIC® user guide: Listening & reading*. Retrieved from https://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: and sex and drugs and rock 'n' roll* (4th ed.). London: Sage.

Furukawa, A. (2010). *Eigo tadoku hou: Yasashii hon de hajimereba tsukaeru eigo ha kanarazu mi ni tsuku* [Methods for extensive reading in English: Starting with easy books, develop useful English for sure]. Tokyo: Chuo Seihan.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Grabe, W. (2010). Fluency in reading – Thirty-five years later. *Reading in a Foreign Language, 22*, 71-83. Retrieved from: http://nflrc.hawaii.edu/rfl/April2010/articles/grabe.pdf

Grabe, W. (2011, September). *Extensive Reading: Why aren't we all doing it?* Plenary presentation at the First Extensive Reading World Congress, Kyoto, Japan. Retrieved from http://erfoundation.org/erwc1/course/view.php?id=14 [Video retrieved from http://www.youtube.com/watch?v=qYPYg6mP5_Y]

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading*. New York: Longman.

Judge, P. (2011). Driven to read: Enthusiastic readers in a Japanese high school's extensive reading program. *Reading in a Foreign Language, 23*, 161-186. Retrieved from http://files.eric.ed.gov/fulltext/EJ943535.pdf

Krashen, S. (2004). *The Power of Reading* (2nd ed.). Portsmouth, NH: Heinemann

Macalister, J. (2010). Investigating teacher attitudes to extensive reading practices in higher education: Why isn't everyone doing it? *RELC Journal, 41*, 59-75. doi: 10.1177/0033688210362609

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Mason, B. (2011). Impressive gains on the TOEIC after one year of comprehensible input, with no output or grammar study. *The International Journal of Foreign Language Teaching, 7*(1), 1-5. Retrieved from http://www.benikomason.net/articles/Mason_Tanaka_IJFLT_11-11.pdf

Mason, B., & Krashen, S. (1997). Extensive reading in English as a foreign language. *System, 25*(1), 91-102.

McDonald, K., & Malarcher, C. (2008). *Extensive Reading for Academic Success, Advanced D*. Tokyo : Compass Publishing.

McLean, S. (2014). Evaluation of the cognitive and affective advantages of the *Foundations Reading Library* series. *Journal of Extensive Reading, 2*, 1-12. Retrieved from http://jalt-publications.org/access/index.php/JER/article/view/864/61

Nation, I. S. P. (2008). *Teaching ESL/EFL Reading and Writing*. New York: Routledge.

Nishizawa, H., Yoshioka, T., & Fukada, M. (2010). The impact of a 4-year extensive reading program. In A. M. Stoke (Ed.), *JALT2009 Conference Proceedings*. Tokyo: JALT. Retrieved from http://jalt-publications.org/recentpdf/proceedings/2009/E035.pdf

Nuttall, C. (2005). *Teaching reading skills in a foreign language* (3rd ed.). Oxford: Macmillan.

O'Neill, B. (2012). Investigating the effects of Extensive Reading on TOEIC® reading section scores. *Extensive Reading World Congress Proceedings, 1*, 30-33. Retrieved from http://erfoundation.org/proceedings/erwc1-ONeill.pdf

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language, 18*, 1-28. Retrieved from http://nflrc.hawaii.edu/rfl/april2006/pigada/pigada.pdf

Renandya, W. A. (2007). The power of extensive reading. *RELC Journal, 38*, 133-149. doi: 10.1177/0033688207079578

Robb, T., & Kano, M. (2013). Effective extensive reading outside the classroom: A large scale experiment. *Reading in a Foreign Language, 25*, 234–247. Retrieved from http://www.nflrc.hawaii.edu/rfl/October2013/articles/

robb.pdf

Robb, T., & Susser, B. (1989). Extensive reading vs. skills building in an EFL context. *Reading in a Foreign Language*, *5*, 239-251. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl52robb.pdf

Ryall, J. (2013, January). Japan firms see importance of speaking in tongues. *Duetche Welle*. Retrieved from http://dw.de/p/17OwA

Stoeckel, T., Reagan, N., & Hann, F. (2012). Extensive reading quizzes and reading attitudes. *TESOL Quarterly, 46*, 187-198. doi: 10.1002/tesq.10

Storey, C., Gibson, K., & Williamson, R. (2006). Can extensive reading boost TOEIC scores? In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2005 Conference Proceedings*. Tokyo: JALT. Retrieved from http://jalt-publications.org/archive/proceedings/2005/E034.pdf

Susser, B., & Robb, T. (1990). EFL extensive reading instruction: Research and procedure. *JALT Journal*, *12*(2), 161-185. Retrieved from http://www.cc.kyoto-su.ac.jp/~trobb/sussrobb.html

Takahashi, J. (2012). An overview of the issues on incorporating the TOEIC test into the university English curricula in Japan. *Tama University Global Studies Department Bulletin*, *4*(3), 127-138. Retrieved from: http://repository.tama.ac.jp/modules/xoonips/detail.php?id=AA12419269-20120331-1010

Takase, A. (2008). The two most critical tips for a successful extensive reading program. *Kinki University English Journal, 1*, 119-136. Retrieved from http://jairo.nii.ac.jp/0066/00000519

Tokunaga, M. (2008). Students' assumptions for TOEIC classes. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT 2007 Conference Proceedings*. Tokyo: JALT.

Yamashita, J. (2008). Extensive reading and development of different aspects of L2 proficiency. *System, 36*, 661-672. doi:10.1016/j.system.2008.04.003