Timed reading: Increasing reading speed and fluency

Andrew Atkins
Kyoto Sangyo University

Reference data:

Atkins, A. (2010) Timed reading: Increasing reading speed and fluency. In A. M. Stoke (Ed.), *JALT2009 Conference Proceedings*. Tokyo: JALT.

An overview of the findings of a mixed-methods investigation into the effectiveness of concurrent timed reading and extensive reading programs in a university context. The data were gathered over a 14-week semester from 5 intact classes, which met twice a week. Quantitative analysis provides information about reading fluency gains, and some reflection on qualitative data sheds further light on the process from the participants' viewpoint.

大学において、timed reading（速読）およびextensive reading（多読）を同時に行うプログラムの有効性を調べるため、混合手法による調査を行い、得られた結果の概略を述べる。このデータは、週2回ずつの授業で、1学期14週間にわたり、5クラスの全員から収集された。定量的な分析により、リーディングの流暢さの進歩に関する情報が得られ、また、定性的なデータについての考察により、このプロセスに参加者の視点から新たな光が当てられた。

Fluency is a largely ignored area of study in the years leading up to university entrance in Japan, but once students arrive in university classrooms the chance exists to redress the issue. Since the 1960s there has been research into Timed Reading (TR), although the name given to it has often been different. Conclusive research into the effectiveness of TR is still unavailable, and this paper is an attempt to move closer to understanding whether as teachers we should be allotting time to the practice.

This paper will examine if regular timed reading leads to gains in reading fluency in intermediate and low ability classes, and also whether the amount of Extensive Reading done has a statistically significant interaction effect with this skill.

There is a distinction between Speed Reading and Timed Reading. Speed Reading, as a Google search of the term confirms, usually refers to a practice that is essentially skimming a text for information and not actually reading every word. Speed reading is usually practiced by native speakers of a language, with reading rates for proficient readers reaching 600 or more words per minute (De Leeuw & De Leeuw, 1965).

TR (Champeau de López, 1993) is the reading of texts of equal length and equal lexical difficulty, regularly over a period of weeks or months. Texts are read against the clock and followed by a set of comprehension questions, which are answered without referring back to the text. The purpose of the activity is to increase reading fluency. It can be assumed that reading speed will increase with practice as will comprehension (Utsu, 2003, 2005), although for some participants, this may not be the case. The power law of practice is a phenomenon that applies to practiced skills of this kind (Newell & Rosenbloom, 1981), where gains at the beginning of a study will be pronounced, but will slowly level off to a stage where the participants become more skillful, and gains are almost unnoticeable (Logan, 1992).

Definitions of reading fluency are many and varied, but in a meta-analysis of definitions from first language (L1) studies, Wolf and Katzir-Cohen (2001) arrived at an amalgamated definition. They state that "reading fluency refers to a level of accuracy and rate where decoding is relatively effortless; … and where attention can be allocated to comprehension" (p. 219). The key constituents of this definition are *accuracy*, *rate* and *comprehension*. Following on from this however, almost every study purporting to examine reading fluency uses reading *rate* as the dependant variable, ignoring *accuracy* and *comprehension* in analyses. This raises issues with the validity of previous studies in both L1 and L2 contexts.

Missing data

In almost any kind of longitudinal study researchers will encounter missing data for some or many of the participants. In the past, the most common method used for dealing with the issue has been to delete these cases from the data and analysis. This causes problems with sample sizes for analysis, often excluding the most interesting cases from the study and potentially distorting the results. Another option for dealing with the problem is to impute the values, that is replace a missing value with an estimate of what it would have been had it been measured. Until recently, with the development of specialized computer software programs, many methods of imputation, although simple to implement, were problematic and lacked a sound statistical base (Darmawan, 2002). However the freeware program NORM (Schafer, 1997) provides a means of imputing missing values using a method of data augmentation called multiple imputation. It is not as good as having the real data, but is an improvement on deleting cases and superior to inserting group means or using only the Expectation Maximization (EM) algorithm to generate values (Allison, 2001). The scope of this paper does not provide space for a full discussion of the procedure, but further details can be found in Allison (2000, 2001).

Literature review

Recent attempts to assess the effectiveness of L2 TR include studies by Chung and Nation (2006), Crawford (2008), and Utsu (2004, 2005). Chung and Nation (2006) suggested that there was no established way to measure reading gains, and they explored three similar methods to assess reading gains using only percentage increase in reading rate. Utsu's (2004, 2005) studies also used reading rate as one dependant variable and then separately used comprehension scores as another dependant variable. These studies only looked at percentage increases in both variables, but found that both rate and comprehension improved in both studies. Crawford's (2008) study used a more powerful and valid means of measuring reading rate, repeated measures ANOVA, and although he also mentioned comprehension, he effectively excluded it from analyses. Crawford also suggested there were validity issues in the study, such as the measurement of times, which appeared to be inaccurate.

The most serious problems with the few studies undertaken to examine the effectiveness of TR were the methods of analyses and the dependant variable used. Crawford's (2008) use of repeated measures ANOVA appeared to be the most appropriate method used so far, and the study provided some positive support for the use of TR. The dependant variable in the study however was essentially reading rate, and this falls short of measuring reading fluency gains. It is possible to read quickly without understanding, and it is therefore prudent to include both reading rate and comprehension in an analysis, as this makes it a more valid means of assessing gains in reading fluency.

Outside of L2 and even L1 studies, another more powerful and robust method for analyzing longitudinal studies exists. Latent variable growth curve modeling provides a stronger method for analyzing a TR study that removes any need for dealing with missing data as it is accounted for by the model, and it allows researchers to compare cases at the group and individual level. Unfortunately, the population size needs to be greater than was available for this study (Duncan, Duncan, & Stryker, 2006).

Statement of hypotheses

The three main hypotheses that this study set out to find support for are:

1. TR leads to improvements in reading fluency skill.
2. Improvements in reading fluency skill will be related to the number of graded readers read.
3. For groups with less vocabulary knowledge, reading fluency gains will be less, but still greater than those who do the readings without the time pressures.

Methods

Participants

Five intact classes of Japanese university students ($n = 101$) took part in the study. The participants were all first year students taking a required English course at a private university in western Japan. The classes were streamed into five levels by the results obtained on a proficiency test created by the university. Level 5 is the highest ability level and level 1 is the least proficient. Two of the classes (class D and class E) in this study were from level 2 ($n = 17$, $n =17$), two of the classes (class B and class C) were from level 4 ($n = 22$, $n = 23$), and the final class (class A) was from level 5 ($n = 22$). I taught all of the classes, thus avoiding any teacher differences.

The level 4 and level 5 classes did TR as part of their regular twice-weekly lessons over a period of 10 weeks. Class E did TR in one of their weekly lessons for a period of 12 weeks. The remaining level 2 class (class D) acted as a comparison group. They did the first and twelfth reading of the series as TR, but reading 2 to reading 11 were studied without time constraints to assess whether the timing had any effect on performance.

Materials

The textbook used for the reading practice was *Reading for Speed and Fluency, Book 1* by Nation and Malarcher (2007). The book was written for L2 learners, using a controlled vocabulary load, and consists of 40 readings each with 300 words, followed by five comprehension questions. See (Atkins, 2009) for a more detailed review.

Procedures

In the second week of their first semester at university, the students were introduced to the textbook and were told in detail about the aims of the course. They were told in Japanese not to skim, but to try and read the passages fluently and make every effort to understand what they were reading. 80% to 100% was set as an objective for comprehension, and in addition to this, students were told that a composite score would be used to check their progress. The composite score was calculated by dividing the total time taken to read a 300-word passage in seconds by the raw score on the 5-point quiz. See Table 1 for some examples. Students were aiming to lower their score over the course of study. Understanding of the procedure and scoring was checked with the students and they showed that they had understood. The scores were not part of students' class grades, and after the objectives of TR had been explained, all students were willing to participate in the study.

Table 1. Examples of scoring system used

	Time (seconds)	Comprehension score	Composite score
Example 1	100	5	20
Example 2	60	5	12
Example 3	60	1	60

Before the first reading was done the students were given the first five levels of the Vocabulary Size Test (Nation & Beglar, 2007) to assess their knowledge of the first 5,000 word families of English. This was done to predict the likely vocabulary coverage of the texts to be used. The scores for the Vocabulary Size Test were compared to the output for the texts obtained from the RANGE program (Nation & Heatley, 2002) and I decided that the students would in most cases have sufficient vocabulary coverage to be able to perform the tasks.

The first reading was done after the explanation and understanding of the procedure had been confirmed with the students. All students performed the task correctly. Subsequent readings were undertaken at the start of each lesson. The level 4 and 5 classes studied two passages a week (one every class), but the level 2 classes could only do one reading a week, because they had a computer-based class for one of their twice-weekly sessions. For students in the treatment groups, each reading was started when they received the signal from the teacher, and when they finished they recorded their time from a large digital chronograph displayed on a projector or computer screen. Once they had noted their time they turned the page and without referring back to the text answered the questions. I then went through the answers with them and they wrote down their comprehension scores. Students then calculated their composite scores and wrote them alongside the score for the quiz in their books. For the comparison group the procedure was the same for the first and twelfth reading, however I removed the time constraint for the other passages.

The time used in class decreased as the semester progressed, and by the tenth reading less than 10 minutes was devoted to the activity in all of the treatment groups, and about 12 minutes were used in the comparison group. The remaining time in the lessons was devoted to other tasks that mainly focused on speaking and listening.

Motivation was generally high and there was a competitive atmosphere in the level 4 classes, with most students comparing scores with their peers after each reading. In the last two weeks of term there appeared to be less effort, perhaps due to tiredness from assignment writing and exam preparation.

The students all took part in the extensive reading program at the university and were encouraged to read books for part of their grade. If a student read 5 books in the semester they were neither penalized nor rewarded for their efforts. Less than 5 books read in the semester meant they would lose 1% of their grade for each book they failed to read. More than 5 books resulted in a bonus of 1% of their grade for each book read and there was no upper limit set as to how many books could be read.

Analysis and results

Missing data

As discussed above, there were some missing data in the study, as is the case in most longitudinal studies of this nature. The missing data for each of the treatment groups are shown in Table 2.

All of the missing data were imputed using the NORM program (Schafer, 1997) and therefore it was unnecessary to delete any cases or variables from the study. The multiple imputations were performed by first generating parameter estimates using the EM algorithm to generate a covariance matrix, and then using the estimated parameters as starting values to perform the multiple imputations, imputing values after every 1,000 replacement situations, and generating the augmented data used in the model.

Table 1 shows the amount of missing data for the treatment groups showing complete and incomplete for each category. For example, for class A, only 10% of the *variables* (readings) have complete data. This means that for the 20 readings measured in class, only two were fully attended. 40.9% of *cases* (participants) have complete data, and this means that 59.1% of participants were absent for one or more reading. 90.9% of *values* (individual measurements) were complete, meaning that of the maximum possible 440 *values* that could have been measured if all students in the class had done every reading, 40 were missing.

The percentage of missing data *values* falls well within the acceptable maximum of 20% recommended by Little and Rubin (2002). Had incomplete cases been deleted from the analysis as has been traditional, more than half the total participants' data would have been lost, biasing the results by using only data from those students with perfect attendance.

Dependant variable

As TR is seen as a means to increase fluency (Nation, 2005), it's obvious that the dependant variable in any study attempting to assess fluency should be a measure of fluency. Fluency is a latent variable and much more than speed or rate. Estimating an object's density by measuring only its weight would not be appropriate, and measuring fluency by only speed appears to be inappropriate too. Speed and fluency will correlate, but this does not mean they are equal. Some of the participants will have focused more on speed than comprehension, while others will have done the reverse.

Table 2. Missing data percentages for treatment groups

	Variables		Cases		Values	
Class^level	Complete	Incomplete	Complete	Incomplete	Complete	Incomplete
Class A^L5 (*n*=22)	10	90	40.9	59.1	90.9	9.1
Class B^L4 (*n*=22)	40	60	36.4	63.6	93	7
Class C^L4 (*n*=23)	30	70	37.5	62.5	95	5
Class E^L2 (*n*=17)	50	50	58.8	41.2	96.1	3.9

A TR study, therefore, should also account for errors made on the comprehension questions. Time taken to read a passage and items answered correctly on the comprehension tests will in some cases be a trade off, but research from L1 studies has actually found that the faster someone reads, the better their comprehension, supposedly due to working memory constraints (Breznitz, 1987, cited in Breznitz, 2006).

For the purposes of the study discussed here, a composite score made from time taken to read a text in seconds, divided by the raw score on the multiple choice comprehension test was used as the dependant variable. In other words, the score for a reading was equal to the number of seconds it took a student to earn one point on the comprehension test. This however, is not without its problems, and further assessment of the validity of the variable is necessary.

Descriptive statistics

Once missing data had been augmented the descriptive statistics for the dataset were calculated. The statistics for the treatment groups' composite reading scores are shown in Table 3. The data have been divided into stages of four readings each; this was chosen instead of using the raw scores for all 20 (or 12 for the level 2 class) readings because there were some relatively large variations between readings and the stage mean provides a more useful and stable view of skill in a two-week period (or four-week period for the level 2 class). The variations between the 20 readings also made it impossible for SPSS to compute a solution using repeated measures ANOVA when all readings were used.

It can be seen by the decrease in the mean composite reading scores that in the first four stages of the study that there were continued improvements for all classes. However, for classes A, B, and C the reading scores increased in stage 5, indicating a decrease in performance. This decrease however was only caused by the results of the performance on one reading. Even though for the reading in question the average reading time remained consistent with other readings, the average score on the comprehension questions was almost one point out of five lower than for other readings. This appears to have been due to a combination of lack of subject knowledge on the part of the students and one poorly written test item.

Table 2 shows the mean composite scores (M) and the standard deviation (SD) by class for each stage (four readings) of

Table 3. Descriptive statistics for treatment groups' composite reading scores

	Stage 1 (1 to 4)		Stage 2 (5 to 8)		Stage 3 (9 to 12)		Stage 4 (13 to 16)		Stage 5 (17 to 20)	
Class $^{\text{level}}$	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Class A$^{\text{L5 }(n=22)}$	33.97	10.88	29.66	12.33	24.61	10.66	20.79	8.18	23.25	7.20
Class B$^{\text{L4 }(n=22)}$	35.41	7.37	27.07	8.05	23.94	8.56	21.04	7.63	24.77	8.83
Class C$^{\text{L4 }(n=23)}$	34.35	8.39	29.70	7.33	22.78	8.97	20.77	4.74	21.96	6.33
Class E$^{\text{L2 (n=17)}}$	67.08	4.49	59.99	3.23	56.45	6.40				

the study. For classes A, B, and C, Stage 4 has the lowest mean composite score, indicating the most fluent reading occurred in this stage.

Repeated measures ANOVA

After the mean composite scores for each stage had been calculated for each student the data were analyzed using SPSS to perform a repeated measures ANOVA. For each of the treatment groups a separate one-way ANOVA was performed and the results are shown in Table 4. For groups A, B, and C it can be seen that they achieved statistically significant improvements over the course of the treatment ($p < .001$). Group E however did not reach significance ($p = .120$). The strength of association (η^2_p) for groups A, B, and C was very high ($\eta^2_p > .53$). See Brown (2008) for an explanation of why partial eta squared should be used over eta squared in longitudinal studies.

Data for classes A, B, and C were then combined and another repeated measures ANOVA was performed in order to compare changes between stages within the study. A pairwise comparison was used to statistically assess differences in performance between the means of stages with all other stages in the repeated measures ANOVA. As could be seen in the descriptive statistics, there was a decrease in performance for stage 5, and this caused a non-significant pairwise comparison between stage 3 and stage 5. All other pairwise comparisons were significant ($p < .001$).

The same three groups were further checked at the group level for significant interactions with a number of independent variables that had been recorded. These included, gender, length of residence overseas, TOEIC scores, vocabulary size, and most importantly for this study, the number of graded readers read over the course of the study. There were no significant interactions with any of the variables, meaning that none had a statistically significant effect on performance in TR. However, for class A the interaction with the number of graded readers read was approaching significance, and therefore we can start to speculate that the more graded readers read the greater improvements in TR will be. This however is ambiguous, as the number of books read could just be an indicator of motivational engagement.

Table 4. Repeated measures ANOVAs by treatment group

Class $^{(level)}$	SS	Df	MS	F	p	η^2_p
Class A$^{L5\ (n=22)}$	1993.33^g	2.331	855.24	17.96	<.001	.60
Class B$^{L4\ (n=22)}$	1930.26	4	482.57	23.93	<.001	.69
Class C$^{L4\ (n=23)}$	2285.42	4	571.36	16.98	<.001	.53
Class E$^{L2\ (n=17)}$	772.70	2	386.35	2.47	.120	.26

^{g} Greenhouse-Geisser correction applied because sphericity assumption was violated.

Counterfactual comparison

Class D acted as a control (counterfactual) group, i.e. they performed the same treatments as class E, but without the time constraints for readings 2 to 11. Ross (1998) suggests that *counterfactuals* "would be expected not to gain at the same rate as the recipients of the program intervention" (p. 37). Class D, because they had no time pressures, spent more time on task than class E. However as can be seen in Table 5 the gain in composite reading score between reading 1 and reading 12 is only slightly higher for class E, and the effect of the time constraint is inconclusive. Class E showed least knowledge on the Vocabulary Size Test, and were chosen as a treatment group over class D because they had the lowest composite reading score in the first reading.

Table 5. Change in scores between reading 1 and reading 12

	Reading 1		Reading 12		Change	
Class level	M	SD	M	SD	M	SD
Class D L2 (*n*=17)	51.82	20.53	37.42	14.18	14.41	23.08
Class E L2 (*n*=17)	59.53	14.80	43.43	9.38	16.10	15.63

Qualitative data

Anonymous qualitative survey data was gathered using open-ended questions at the end of the study. Comments from students were almost all positive. The most common reasons why students felt some readings were more difficult than others were subject familiarity and unknown proper nouns. Two students in the study said that they found TR a little stressful because they were unable to read as fast as others. One student thought TR was a waste of time. The vast majority of students however thought it was beneficial for their reading fluency, with some students asking for further readings to do at home after the study was completed.

Conclusions

The results of this study are not conclusive, but with regard to the first hypothesis, they provide further support that TR leads to gains in reading fluency especially when vocabulary coverage is sufficient. Had there been greater control of vocabulary load, schema, and of test item difficulty, the results may have given more definitive support to the effectiveness of TR in Japanese university classrooms.

There was no significant statistical relationship between improvement in reading fluency and the number of graded readers read. Therefore, we cannot say with any certainty that reading graded readers has an effect on performance in TR. Using the number of books read as the unit of measurement may have been the reason why there was no interaction between TR gains and extensive reading. The books that class A read were generally much longer than the books that classes B and C read because they were at a higher level, and had *words read* been used, the interaction may have been significant for class A. This however is speculation.

Differences between class D and class E were small and inconclusive, however the treatment group did make more gains even though they started from a lower position initially. Had the treatment been twice weekly, more noticeable gains may have been apparent, but this needs to be tested with further research.

Further research

Vocabulary coverage needs to be more stringently controlled as does test item difficulty in further studies. It may be necessary to devise a means of weighting questions to make comparison between readings more valid. Subject familiarity is an issue, and I feel that the only way to perform a more conclusive study is to write passages and questions with greater constraints in these areas.

There may be other variables not assessed in this study that have an interaction with TR performance, and these could include a number of individual difference variables. Repeated measures ANOVA seems to fall short of what is necessary for a study of this kind, and latent variable growth curve modeling may offer a better solution.

The optimum length of a TR passage is still untested in research and needs to be identified. Whether online delivery is as effective as paper is also an issue in need of research. The dependant variable used in this study also needs to be validated using statistical means.

Bio data

Andrew Atkins is a lecturer at Kyoto Sangyo University, and a member of the sixth doctoral cohort at Temple University, Osaka Campus. He is currently interested in EFL fluency improvement, and is coordinator of the JALT Study Abroad SIG. <andrew556@gmail.com>

References

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research*, *28*, 301-309.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Atkins, A. (2009). Reading for speed and fluency 1. *The Language Teacher*, *33*(7), 43-44.

Breznitz, Z. (2006). *Fluency in reading*. Mahwah, NJ: Erlbaum.

Brown, J. D. (2008). Statistics corner. Questions and answers about language testing statistics: Effect size and eta squared. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *12*(2), 36-41.

Champeau de López, C. L. (1993). Developing reading speed**.** *English Teaching Forum*, *31***(1), 50-51.**

Chung, M., & Nation, I. S. P. (2006). The effect of a speed reading course. *English Teaching*, *61*(4), 181-204.

Crawford, M. J. (2008). Increasing reading rate with timed reading. *The Language Teacher*, *32*(2), 3-7.

Darmawan, I. G. N. (2002). NORM software review: Handling missing values with multiple imputation methods. *Evaluation Journal of Australasia*, *2*(1), 20-24.

De Leeuw, E., & De Leeuw, M. (1965). *Read better, read faster*. London: Penguin.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modelling: Concepts, issues, and applications*. Mahwah, NJ: Erlbaum.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing* data (2nd ed.). New York: Wiley.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 883-914.

Nation, I. S. P. (2005). Reading faster. *PASAA*, 36, 21-37. Retrieved September 23, 2009, from http://www.victoria.ac.nz/lals/staff/Publications/paul-nation/2005-Reading-faster.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13.

Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, I. S. P., & Malarcher, C. (2007). *Reading for speed and fluency, Book 1*. Seoul: Compass.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Ross, S. (1998). *Measuring gain in language programs: Theory and research.* Sydney: Macquarie University, National Centre for English Teaching and Resource.

Schafer, J. L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model (Version 2) [Computer software]. Retrieved from http://www.stat.psu.edu/~jls/misoftwa.html#win

Utsu, M. (2004). Timed readings *no riyou to sono kouka* [Timed readings and their effects on students]. *Bulletin of Yonezawa Women's College of Yamagata Prefecture*, 39, 31-37.

Utsu, M. (2005). Timed readings *no riyou to sono kouka 2* [Timed Readings and their effects on students (Part II)]. *Bulletin of Yonezawa Women's College of Yamagata Prefecture*, 40, 27-34.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, *5*(3), 211-239.