Student clustering in an ER program

Omar Karlin
Tokai University
Rick Romanko
Wayo Women's University

Reference data:

Karlin, O., & Romanko, R. (2008). Student clustering in an ER program.
In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*. Tokyo: JALT.

This study examined the gains in student affect, vocabulary, and reading fluency for 110 university students in an extensive reading program in Japan. It was important to measure all of these dimensions within a single study and a single teaching methodology, so gains could be appropriately compared against each other. The adopted teaching methodology was a communicative one which stressed a number of in-class activities, with out-of-class reading. Students were measured over the course of a semester, and paired-sample t-tests were conducted using pre- and post-test scores on six variables. Results indicated that affect increased a great deal, while fluency increased minimally, and vocabulary did not increase at all. Students were also clustered into two groups, higher-level and lower-level, to determine if they differed at all in their learning rates. Paired sample t-tests indicated that the lower-level students gained more in terms of fluency than the higher-level students.

本研究では、日本の大学の英語多読プログラムの受講者110名の学生を対象に、学生の情緒的反応、語彙、読みの流暢さに関する進歩状況について調査した。進歩を互いに適切に比較することが可能になるように、これらの特質はすべて単一の研究及び単一の教授法のなかで測定することが重要であった。採用された教授法は、授業内における多数のアクティビティに重点をおき、併せてリーディングを課外に行うコミュニカティブ・アプローチであった。学生たちは半期の授業に渡って測定された。6つの変数に関して受講前のテスト・スコアと受講後のテスト・スコアを用いて、対標本t検定が実施された。結果は、情緒的反応の飛躍的な増加を示していた。一方で、読みの流暢さの増加は最小限にとどまり、語彙に関しては全く増加が見られなかった。さらに、学習率に何らかの違いがあるかどうかを決定するため、学生たちはレベルの高い方のグループと低い方のグループの二つの集落に分けられた。対標本t検定の結果は、レベルの低い方の学生たちが、レベルの高い方の学生たちよりも、流暢さの点で進歩の度合いが高かったことを示した。

For English students, especially those in an EFL context where access to an L1 community is constrained and viable language input is minimal, there are a number of benefits realized with extensive reading. One of the central tenets of extensive reading (ER) is the potential boon to affective dimensions, such as motivation, perceived competence, and enjoyment. With respect to research done in Japan, a number of studies have shown ER-related benefits to affect, specifically, university student motivation (Robb & Susser, 1989), high school student motivation (Hashimoto et al., 1998), and even teacher motivation (Takase, 2006).

While not enjoying as much of a consensus in support, vocabulary-learning is another oft-cited benefit. There has been some disagreement over the amount of vocabulary that is actually learned through ER, from more optimistic proponents (Mason & Krashen, 1997) to more tempered views (Waring and Takaki, 2003). However, it is worth noting that even though some have found fault with the lack of rigor in the more optimistic studies (Nation, 1999; Waring & Takaki, 2003), critics of overzealous research have contended that ER can do a lot to help strengthen existing vocabulary knowledge (Waring and Takaki, 2003p. 154).

One final benefit to be mentioned in this brief outline is the effect ER has upon reading fluency. The nature of ER, in which repeated exposure to letters, words, and even texts is maximized, is ideally suited for creating reading automaticity (Logan, 1997) and this has been borne out in the research (Taguchi, Takayasu-Maass, & Gorsuch, 2004). Suffice it to say, research has been supportive, in varying degrees, of affective, lexical, and reading fluency benefits catalyzed by ER.

However, one area that has been lacking is the examination of these multiple variables within a single study and a single research methodology. It is difficult to take research at face value when ER practitioners often rely on vastly different teaching approaches, with some adhering to a hands-off approach in which students are encouraged to read within the class and not saddled with reports, assignments, or other burdensome activities that might sap the intrinsic development of a reading habit (Day & Bamford, 2002). While others feel that a more comprehensive approach that integrates skills and recycles concepts through a variety of assignments is to the ultimate benefit of students (Hunt & Beglar, 2005).

If one study claims a vocabulary result with the former methodology, and another claims a motivational benefit with the latter methodology, questions should be raised as to whether the results are commensurate. In fact, conditions that spur vocabulary development for example, might rely on rigid recycling of vocabulary through book reports and vocabulary-isolating activities, while a study that is intent on fostering affective gains, may intentionally minimize such conditions, thereby making these results problematic when cited as uniform benefits of ER. It is crucial to view all of these potential benefits in a unified research design, not only to eliminate potential contradictions, but also to give insight as to the degree that certain phenomenon are occurring in relation to each other.

Another area of ER that may be under-researched, at least in relation to the aforementioned benefits of affect, vocabulary, and fluency, is how these benefits are realized by the different students within a class. Clearly, not every learner is the same, and the benefits of ER weigh upon different learners in different ways. In an interesting study on demotivation factors that impede poorly motivated ER students, Takase outlined the obstacles separating the highly-motivated from the poorly-motivated (Takase, 2003). In another study, Mori identified several predictors that manifested in different reading intensities in ER students, implicitly distinguishing between different student types (Mori, 2004).

While these studies examined the motivation and reading intensity of different types of students in an ER class, they did not focus on linguistic variables such as vocabulary knowledge and reading fluency. This study proposes to identify class-wide benefits of ER, identify the different types of students in an ER class, and contrast the benefits realized by these different types of students. It should be noted that the teaching methodology used in this study had a heavy focus on skill integration within the class through communicative activities, and independent reading outside of class.

Research questions

Examining the three dimensions of affect, vocabulary, and reading fluency in conjunction provides a more comprehensive view of the effects of ER and its overall benefits for students. Examining various groups of students within an ER program may provide insights as to the best utility of ER courses within an established curriculum. This study will attempt to answer the following three hypotheses.

Hypotheses

1. Participants, as a whole, will see a significant improvement in affect and reading fluency, but not in vocabulary knowledge.
2. Student affect will show more substantial gains than fluency, and both will show more significant gains than vocabulary (if there are any).
3. Clusters of students will experience different degrees of success in the ER program, as evidenced by post-test scores.

Research design

Participants

Participants in this study included 116 first and second year students at a national university in Tokyo, Japan. All participants were non-English majors enrolled in a compulsory English reading course in which the medium of instruction was English. Three students had incomplete data, and three students were statistical outliers, resulting in all six being dropped from the study. The final number of participants was 110 (85 male and 25 female). Participants came from three separate classes which were assembled based on their major and/or faculty, and all classes were of approximately the same proficiency.

Procedure

The first meeting of the course was used as an orientation class in which students learned about the course syllabus, rules, and philosophy of the class: *The best way to become better readers is through reading, reading, and more reading!*

It was explained to students that they would be expected to read at least ten graded readers of their choice (about a book a week) during the semester. The reading would be done on their own time outside of class. During the second meeting, a series of questionnaires and tests was administered to students in order to measure their affect, vocabulary, and reading fluency. Similar tests were administered again during the final meeting of the semester.

Affect

With regard to affect, a 13-item questionnaire was created based upon self-efficacy principles, which are essentially an evaluation of self (Dornyei, 2005). The questionnaire was arranged in two sections. The first section (á = .815), comprised of 6 items, focused on reading ability, while the second section (á = .894), comprised of 7 items, focused on overall English ability. Students answered items on a 4-point Likert scale that ranged from strongly negative to strongly affirmative. Simple and easy-to-understand English was used for all of the questionnaire items.

It was believed that if questionnaire items were written at an appropriate level of English, it would not hinder students' comprehension. Two other native English professors also looked at the items to form a consensus that the level of English was appropriate (Brown, 2001). For the statistical analysis, each section of the questionnaire was totalled to form a composite score, one relating to reading ability and the other relating to overall English ability.

Vocabulary

For vocabulary assessment, students completed Nation's 2000-word level and 3000-word level productive and receptive vocabulary tests (Nation & Laufer, 1999). There was a noticeable *floor effect* involving the 3000-word level tests (i.e. all of the students were scoring poorly on it), so it was dropped from the statistical analysis. In addition to the 2000-word level tests, students were given a 2000-word level *Yes/No* perception test in which they were asked to estimate for themselves how many words they did not know.

The intent of this test was to provide some insight as to the students' *perceptions* of their vocabulary (not their *actual* vocabulary). Since it was a test of words they did not know, a lower score actually meant an improving vocabulary (this is important to remember when viewing the statistical results). In all, three tests were included in the statistical analysis, Nation's 2000-word level productive and receptive tests (assessing actual vocabulary knowledge) and a 2000-word level *Yes/No* perception test (assessing perceived vocabulary knowledge).

However, the 2000-word level *Yes/No* perception test was not considered a reliable indicator of vocabulary knowledge because it lacked rigor. Rather, it was considered an affective test, measuring students' perceptions of their own development and competence.

Reading fluency

Finally, with regard to reading fluency, students were asked to read a passage of text at a speed that was comfortable for them. The number of words in the passage was divided by the amount of time taken to read the passage, to establish a *words-per-minute* (WPM) score. A comprehension test of 5 questions was administered after students finished reading to ensure they abided by the rules of the activity and did not recklessly speed-read (Nuttall, 1996). A WPM score was used in the statistical analysis to represent reading fluency.

Statistical analysis

Six paired-sample t-tests were conducted using data from 110 students. Each paired-sample t-test was based on a pre- and post-test variable. Respectively, the six variables were a 2000-word level *Yes/No* perceived vocabulary test, 2000-word level production and reception vocabulary tests, reading ability and overall English ability questionnaire results, and words-per-minute scores (WPM). For both the initial class-wide analysis and the secondary student cluster analysis, a one-tailed hypothesis was selected since it was thought that affect and fluency would significantly improve, while vocabulary would not. Also, post-hoc correction methods were conducted using Holm's sequential procedure. It was thought that Holm's procedure would have the statistical power to avoid Type 1 errors, yet be flexible enough to also prevent Type II errors, as is evident in the second analysis involving different clusters of students (Holm, 1979).

Following the class-wide paired-sample t-tests, a hierarchical cluster analysis was conducted in order to determine the most appropriate number of clusters in which to divide the students. The selected cluster variables were the 2000-word level perceived vocabulary pre-test, the 2000-word level production and reception pre-tests, the reading ability and overall English ability questionnaire pre-tests, and the WPM pre-tests. The cluster method selected was centroid clustering, the measure was the interval of the squared Euclidian distance, and the values were transformed into z-scores. Results indicated that a two-cluster solution would be the most appropriate. A subsequent k-means cluster analysis was conducted with two clusters specified as a solution.

Following the cluster analysis, paired-sample t-tests were conducted again for each cluster, using pre- and post-test scores. The same six paired-sample variables that were used in the class-wide t-tests were again used for the cluster t-tests. Again, Holm's sequential procedure was used as a post-hoc correction method.

Results

The descriptive statistics and correlations for the class-wide paired-sample t-tests can be found in Table 1. Of note, the value for the 2000-word perception pre- and post-tests represent the number of *unknown* words, hence the decreasing number in the post-test. Also, correlations between the paired-sample variables were very strong, which was to be expected since each paired-sample tested the same construct through a pre- and post-test.

The results of the paired-sample t-tests for all of the participants can be found in Table 2. On average, participants experienced significantly higher scores on the post-tests for the three affective measures (pair 1, pair 4, and pair 5) when compared to their pre-test scores. Also of note, the effect size for these three pairs was considerable. With regard to reading fluency, participants experienced a significant increase on the post-test (pair 6). However, it should be noted that the effect size for this increase in reading fluency was rather small. Finally, participants did not experience any significant increases in terms of actual vocabulary knowledge, as noted on the 2000-word production and reception post-tests (pairs 2 and 3).

After conducting class-wide paired-sample t-tests, a cluster analysis was performed in order to segment the class into different student clusters. The results indicated that a two-cluster solution was most appropriate for the available data. Results of the cluster analysis are summarized in Table 3. The larger of the two clusters, cluster one (N=75), scored higher on all of the variables.

Cluster one had better vocabulary knowledge, higher perceived reading and English ability (the questionnaires), and read more words-per-minute. The only area in which cluster one had a lower score than cluster two was in the 2000-word level perception vocabulary test, which again indicates a superior perception of ability because this score indicates *unknown* words. As a result, we characterized these students as “higher-level” (cluster one) and “lower-level” (cluster two).

The descriptive statistics and correlations for cluster one’s paired-sample t-tests (N=75) can be found in Table 4. Of note, correlations between the paired-sample variables were generally weaker for cluster one than they were for the class-wide paired-sample t-tests (Table 1).

The results of the paired-sample t-tests for cluster one can be found in Table 5. Again, affective measures (pair 1, pair 4, and pair 5) were significantly higher on the post-tests than on the pre-tests. Effect sizes were also considerable. Of note, reading fluency (pair 6) did not achieve significance, and after using the Holm’s sequential procedure post-hoc test, it was not as close to significance as it initially appeared (with the significance threshold for pair 6 settling at 0.017, well below the actual 0.058 indicated). Finally, cluster one participants did not experience any significant increases in terms of actual vocabulary knowledge, as noted on the 2000-word production and reception post-tests.

The descriptive statistics and correlations for cluster two’s paired-sample t-tests (N=35) can be found in Table 6. Of note, correlations between the paired-sample variables were stronger than for both cluster one (Table 4) and the class-wide sample (Table 1). Also, the variables in pair 1 did not significantly correlate.

Table 1. Descriptive statistics and correlations for class-wide paired-sample t-tests

		Mean	ά.	SE	Corr.	Sig.
Pair 1	2000-word perception pre	137.34	86.79	8.28	.51	.000**
	2000-word perception post	20.84	22.54	2.15		
Pair 2	2000-word production pre	10.77	3.21	.31	.39	.000**
	2000-word production post	10.75	3.26	.31		
Pair 3	2000-word reception pre	24.70	2.66	.25	.61	.000**
	2000-word reception post	24.65	2.79	.27		
Pair 4	Reading ability pre	14.57	2.62	.25	.41	.000**
	Reading ability post	17.17	2.67	.25		
Pair 5	English ability pre	12.34	3.14	.30	.42	.000**
	English ability post	15.94	3.27	.31		
Pair 6	Words-per-minute pre	119.09	35.36	3.37	.54	.000**
	Words-per-minutes post	128.28	33.73	3.22		

Note: ** significant using Holm’s sequential procedure (beginning at 0.05)

Table 2. Class-wide paired-sample t-tests

		Mean	á.	SE	95% CI		t	df	Sig.	r
					lower	upper				
Pair 1	2000-word perception pre 2000-word perception post	116.50	77.82	7.42	101.79	131.21	15.70	109	.000**	.69
Pair 2	2000-word production pre 2000-word production post	.02	3.55	.34	-.65	.69	.05	109	.957	.00
Pair 3	2000-word reception pre 2000-word reception post	.05	2.40	.23	-.41	.50	.20	109	.843	.00
Pair 4	Reading ability pre Reading ability post	-2.60	2.86	.27	-3.14	-2.06	-9.52	109	.000**	.45
Pair 5	English ability pre English ability post	-3.60	3.45	.33	-4.25	-2.95	-10.96	109	.000**	.52
Pair 6	Words-per-minute pre Words-per-minutes post	-9.19	33.12	3.16	-15.45	-2.93	-2.91	109	.004**	.07

Note: ** significant using Holm's sequential procedure (beginning at 0.05)

Table 3. Final cluster centers for cluster analysis

	Cluster	
	1 (75N)	2 (35N)
2000-word perception pre	87.43	244.29
2000-word production pre	11.20	9.83
2000-word reception pre	25.36	23.29
Reading ability pre	14.75	14.20
English ability pre	12.49	12.00
Words-per-minute pre	123.71	109.20

Table 4. Descriptive statistics and correlations for cluster one's paired-sample t-tests

		Mean	á.	SE	Corr.	Sig.
Pair 1	2000-word perception pre 2000-word perception post	87.43 14.39	44.00 15.19	5.08 1.75	.46	.000**
Pair 2	2000-word production pre 2000-word production post	11.20 11.00	2.95 3.26	.34 .38	.28	.014**
Pair 3	2000-word reception pre 2000-word reception post	25.36 25.14	1.50 1.88	.17 .22	.23	.047
Pair 4	Reading ability pre Reading ability post	14.75 17.22	2.53 2.58	.29 .30	.37	.001**
Pair 5	English ability pre English ability post	12.49 16.25	3.15 3.38	.36 .39	.43	.000**
Pair 6	Words-per-minute pre Words-per-minutes post	123.71 131.57	37.76 33.56	4.36 3.88	.51	.000**

Note: ** significant using Holm's sequential procedure (beginning at 0.05)

Table 5. Cluster one paired-sample t-tests

		Mean	á.	SE	95% CI		t	df	Sig.	r
					lower	upper				
Pair 1	2000-word perception pre 2000-word perception post	73.04	39.43	4.55	63.97	82.11	16.04	74	.000**	.77
Pair 2	2000-word production pre 2000-word production post	.20	3.73	.43	-.66	1.06	.47	74	.644	.00
Pair 3	2000-word reception pre 2000-word reception post	.21	2.12	.24	-.27	.70	.87	74	.386	.10
Pair 4	Reading ability pre Reading ability post	-2.48	2.87	.33	-3.14	-1.82	-7.49	74	.000**	.43
Pair 5	English ability pre English ability post	-3.76	3.50	.40	-4.57	-2.95	-9.30	74	.000**	.54
Pair 6	Words-per-minute pre Words-per-minutes post	-7.87	35.40	4.09	-16.01	.28	-1.93	74	.058	.05

** significant using Holm's sequential procedure (beginning at 0.05)

Table 6. Descriptive statistics and correlations for cluster two's paired-sample t-tests

		Mean	á.	SE	Corr.	Sig.
Pair 1	2000-word perception pre 2000-word perception post	244.29 34.66	51.64 28.95	8.73 4.89	.19	.274
Pair 2	2000-word production pre 2000-word production post	9.83 10.20	3.56 3.24	.60 .55	.58	.000**
Pair 3	2000-word reception pre 2000-word reception post	23.29 23.60	3.84 3.95	.65 .67	.72	.000**
Pair 4	Reading ability pre Reading ability post	14.20 17.06	2.81 2.89	.47 .49	.49	.003**
Pair 5	English ability pre English ability post	12.00 15.26	3.15 2.96	.53 .50	.40	.016**
Pair 6	Words-per-minute pre Words-per-minutes post	109.2 121.23	27.51 33.46	4.65 5.66	.60	.000**

** significant using Holm's sequential procedure (beginning at 0.05)

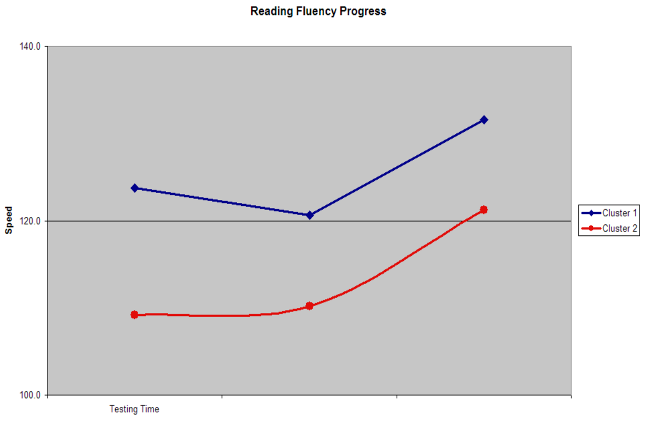

Figure 1. Reading fluency progress for both clusters

Table 7. Cluster two paired-sample t-tests

		Mean	á.	SE	95% CI		t	df	Sig.	r
					lower	upper				
Pair 1	2000-word perception pre 2000-word perception post	209.63	54.19	9.16	191.01	228.24	22.89	34	.000**	.94
Pair 2	2000-word production pre 2000-word production post	-.37	3.13	.53	-1.45	.71	-.70	34	.488	.01
Pair 3	2000-word reception pre 2000-word reception post	-.31	2.92	.49	-1.32	.69	-.64	34	.528	.01
Pair 4	Reading ability pre Reading ability post	-2.86	2.88	.49	-3.85	-1.87	-5.87	34	.000**	.50
Pair 5	English ability pre English ability post	-3.26	3.35	.57	-4.41	-2.10	-5.76	34	.000**	.49
Pair 6	Words-per-minute pre Words-per-minutes post	-12.03	27.88	4.71	-21.61	-2.45	-2.55	34	.015**	.16

** significant using Holm's sequential procedure (beginning at 0.05)

The results of the paired-sample t-tests for cluster two can be found in Table 7. Again, affective measures (pair 1, pair 4, and pair 5) were significantly higher on the post-tests than on the pre-tests. Effect sizes were also considerable. Also of note, reading fluency (pair 6) achieved significance (with the Holm's procedure threshold settling at 0.017). This is a key divergence from cluster one, which did not achieve significance in terms of reading fluency. Finally, cluster two participants did not experience any significant increases in terms of actual vocabulary knowledge, as noted on the 2000-word production and reception post-tests.

The divergence in reading fluency between cluster one and cluster two is illustrated in Figure 1. Reading fluency was assessed at three points during the study. The first and last points on the graph are the same as the information included in Tables 4-7. In addition, a fluency score was assessed in the middle of the semester, for which cluster one had a mean average of 120.67 and cluster two had a mean average of 110.20. It should also be noted that when a one-way ANOVA was conducted between the clusters with regard to reading fluency, there was a significant difference between the clusters when measured at the beginning of the semester, $F\ (1,\ 108) = 4.13,\ p < 0.05$. However, there was not a significant difference between the clusters when measured at the end of the semester, $F\ (1,\ 108) = 2.27,\ p > 0.05$

Discussion

The first hypothesis predicted significant gains in student affect and reading fluency, and as predicted, the class-wide sample saw significant gains in both areas (Tables 1 and 2).

Also as predicted, vocabulary did not significantly improve. The nature of ER, diverging from more traditional methods of instruction, seemed to have a quick and profound impact upon student affect. After six years of grammar-translation exam preparation in junior and senior high school, it is not surprising that students' affect improved so dramatically, considering the emphasis ER places on student enjoyment. However, it would be interesting to see if these gains could be maintained over a longer timeline. It is likely that to permanently change student perceptions of their vocabulary, reading, and English abilities, they will need to experience more than a single semester of ER. If delayed post-tests were conducted a few months after the end of the semester, it is doubtful whether affect scores would remain as high.

With regard to fluency, there was a significant increase, yet the effect size was disappointingly small. Again, a longer study may yield additional insights as to the rate of fluency improvement. It would seem that fluency should eventually level-off as students approach the upper-limits of their capabilities. However, in the case of this study their fluency actually appeared to accelerate as they approached the end of the semester. Also, claims in previously-mentioned research of vocabulary knowledge not improving with ER appear to be valid.

In fact, vocabulary knowledge scores actually *decreased* over the semester, although this was not statistically significant. There may have been a couple of reasons behind this lack of vocabulary development. It is possible that the emphasis on communicative activities during class time reduced the amount of time students could have spent reading. During the study, it was thought that the communicative activities may have complemented the reading by providing multi-disciplinary opportunities for recycling. However, it is certainly possible that students were not actually recycling the vocabulary that they read, and were missing opportunities at recycling through more in-class reading. Another possible reason for this lack of vocabulary gain may have been because students were free to choose graded readers from various publishers, specifically Penguin and Oxford. It is possible that each publisher emphasizes different reading lists, resulting in less repetition of key vocabulary. If students had read from a single publisher, there may have been more vocabulary recycling, and as a result, greater vocabulary acquisition. This is certainly worth exploring in future research.

Taking all three variables into account, Tables 1 and 2 seem to suggest a benefits-hierarchy, with increased student affect as the quickest and most pronounced byproduct of ER, followed by slower and smaller increases in reading fluency, and vocabulary knowledge not improving at all. It is important to remember that these are the results of a communicative teaching approach with reading done outside of class. An alternative teaching methodology may yield different results, and would certainly be a worthy course of study for future researchers.

These findings have pedagogical implications in that single one-off ER courses, offered by many universities in Japan, are likely not enough to help boost students' linguistic abilities. Unfortunately, educational institutions that suggest students enroll in ER classes to assist in preparation for proficiency tests (like TOEFL) may be better served by also offering intensive reading courses and coordinating

ER courses with complementary English courses, ensuring the recycling of concepts, grammar, and vocabulary. In this particular instance, ER seemed to do little more than enhance students' perceptions of their abilities, which may be a very worthy goal in some situations.

The cluster analysis essentially divided the class into a two-thirds and a one-third group, with a "higher-level" cluster of 75 and a "lower-level" cluster of 35. Differences between the clusters were most pronounced in terms of linguistic abilities (i.e. productive vocabulary, receptive vocabulary, and reading fluency), and not as pronounced in affective areas (particularly in their perceptions of their reading and English abilities). Both clusters developed in a similar fashion over the course of the semester, except with regard to fluency. During the study, the "lower-level" cluster narrowed the reading fluency gap with the "higher-level" cluster, suggesting that perhaps this type of ER class may yield more benefits for lower-level students.

To confirm this, additional research would need to be conducted, possibly involving students from different class levels in order to accentuate the differences. In this study, the two clusters began the semester as significantly different groups, but ended the semester as significantly similar. The higher-level cluster's gain did come close to an uncorrected significance threshold of 0.05, but after the post-hoc correction brought the significance level down to 0.017, they were not as close as initially thought. On the other hand, the lower-level cluster just barely achieved statistical significance, coupled with a larger effect size. It is not entirely clear as to why the lower-level students' fluency increased at a greater rate, however there are a few possibilities that could be explored in future research. One possibility is that the wide-spread emphasis on creating interest in ER for lower-level students may have resulted in an accidental neglect of higher-level students.

With publishers increasingly aiming for students at the lower end of the spectrum, with a larger selection of very easy graded readers, it is possible that the availability of more challenging graded readers has not been able to keep pace. It would be interesting to examine publisher title-lists and see exactly how many titles are available at each reading level. Further, there seems to be a great deal of emphasis on making the easier titles more accessible, through more pictures and more vivid colour pagination. Again, it would be interesting to see if more challenging graded readers have been able to keep up with the aesthetic enhancements made to easier graded readers. Future research could examine if there are correlations between available graded reader titles at each level, aesthetic enhancements of graded readers, reading fluency, and total pages read.

Another possible explanation for the difference in reading fluency gains between the clusters may have been that the higher-level cluster was less convinced of ER's painless approach to reading. Especially in Japan, a country where students endure long years of arduous studying, the higher-levels may have been more likely to subscribe to a *no-pain, no-gain* reading philosophy. While lower-levels may have been consumed with the success they were finally feeling in an English class, the higher-levels may have felt that the reading was far too easy for them, especially since they had been successful earlier in their scholastic careers with far more demanding tasks. The lower-level cluster actually read more pages and had a larger increase in reading ability perception than the higher-level cluster, although neither of these differences was large enough to register statistical significance.

One final explanation may be that reading fluency eventually begins to level-off as students reach the ceiling of their abilities, giving an advantage to students farther away from the ceiling. However, with the higher-level cluster only reading at 131 WPM, it is debatable as to whether they were approaching the ceiling of their abilities.

Limitations

The results of this research showed a significant class-wide improvement in reading fluency but the abbreviated duration of the study likely truncated the reading fluency improvement. A longer study, perhaps over an entire year (two semesters or more), may have generated more pronounced reading fluency increases, and may have resulted in an even smaller gap between the two clusters.

Additionally, vocabulary knowledge increases may have become evident over a longer research period, especially since their acquisition is contingent upon recycling. Finally, a longer study with delayed post-tests may have revealed the permanence or impermanence of the student affect increase. As the most positive result in this study, it is crucially important to determine if these affective gains will last.

Also, in order to maintain a healthy sample size for the second part of the analysis involving cluster comparisons, it was thought that introducing a control group for the first class-wide analysis should not be pursued. However, comparing the effect of different ER approaches on student gains in affect, vocabulary, and reading fluency is under-explored and very important. Future research may want to replicate this study, but with a larger sample size that allows for multiple conditions, such as only reading with no supplementary in-class activities.

There were also some problematic testing issues involving the sensitivity of the vocabulary tests, and the frequency of the fluency tests. Nation's 2000-word level productive and receptive tests may have been too broad to measure the small number of new words introduced via graded readers. It is possible that students did in fact learn some new vocabulary which was not represented on the 2000-word level tests.

A more sensitive vocabulary test that isolates key vocabulary targets within the graded readers may yield more positive results. Although, getting access to the word-lists used by publishers is often a closely-guarded industry secret. Additionally, multiple fluency tests at each of the testing periods, averaged to create a mean score, may have yielded more reliable fluency scores. Only testing students once at the beginning, middle, and end of the semester, allows for the possibility of an anomalous bad test that could skew the results. This may have been the case with the higher-level cluster's mid-semester fluency score which actually decreased, shown in Figure 1.

Conclusion

ER has been gaining credibility in Japan as an effective way of boosting student affect, strengthening vocabulary knowledge, and increasing reading fluency. The increasing number of ER studies based in Japan is evidence of its growing acceptance as a legitimate pedagogical approach. What the research community has not yet addressed, however, is how different ER approaches yield benefits in varying degrees, and how different students benefit in different ways.

What this study has attempted to demonstrate is that gains among affect, vocabulary, and fluency are very different when examined within a single teaching framework, and that not all students follow the same learning trajectory in ER classes. By framing these issues within a practical framework, this study attempts to merge some of the established ER theory with practical pedagogical goals. At the very least, hopefully this study will prompt others to challenge these assumptions and provide additional insights as to how ER works in a practical and multi-contextual classroom setting.

Omar Karlin is an Assistant Professor at Tokai University, just outside of Tokyo, Japan. He has completed his doctoral coursework at Temple University Japan and is in the midst of completing his dissertation. His dissertation focus is on the interaction of personality, second language acquisition, and different learning environments.

Rick Romanko is a full-time lecturer at Wayo Women's University in Chiba, Japan. He holds an M.Ed. in TESOL from Temple University. His research interests include developing extensive reading programs and how learning is enhanced by various teaching methodologies. He is also active in materials development.

References

Brown, J., D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Day, R., & Bamford, J. (2002). Top Ten Principles for Teaching Extensive Reading. *Reading in a Foreign Language*.

Dornyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hashimoto, M., Takada, T., Isobe, T., Sakai, N., Ikemura, D., & Yokogawa, H. (1998). Reading shido 12 no approach [Twelve approaches to teaching reading]. *Eigokyoiku [English Education], 47*(2), 42-43.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal Statistics, 6*, 65-70.

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language*.

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly, 13*(2), 123-147.

Mason, B., & Krashen, S. (1997). Extensive Reading in English as a Foreign Language. *System, 25*(1), 91-102.

Mori, S. (2004). Significant Motivational Predictors of the Amount of Reading by EFL Learners in Japan. *RELC Journal: A Journal of Language Teaching and Research in Southeast Asia, 35*(1), 63-81.

Nation, P. (1999). Learning vocabulary in another language. *English Language Institute Occasional Publication 19*.

Nation, P., & Laufer, B. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33-51.

Nuttall, C. (1996). *Teaching Reading Skills in a Foreign Language. New Edition*. Portsmouth, NH: Heinemann.

Robb, T. N., & Susser, B. (1989). Extensive Reading vs. Skills Building in an EFL Context. *Reading in a Foreign Language, 5*(2), 239-251.

Taguchi, E., Takayasu-Maass, M., & Gorsuch, G. J. (2004). Developing Reading Fluency in EFL: How Assisted Repeated Reading and Extensive Reading Affect Fluency Development. *Reading in a Foreign Language, 16*(2), 70-96.

Takase, A. (2003). *Effects of eliminating some demotivating factors in reading English extensively.* Paper presented at the JALT2003, Shizuoka.

Takase, A. (2006). *Teachers motivated by students' extensive reading: A case study of teachers' motivation to start reading English books*. Paper presented at the JALT 2005 Conference Proceedings, Tokyo.

Waring, R., & Takaki, M. (2003). At What Rate Do Learners Learn and Retain New Vocabulary from Reading a Graded Reader? *Reading in a Foreign Language, 15*(2).