An Automated Reading Test Bank for Assessing Student Reading Ability

Tae Kudo (School of Science and Technology)
Kym Jolley (School of Science and Technology)
Sei Sumi (School of Science and Technology)
Joshua Wilson (School of Science and Technology)
Laura Huston (School of Science and Technology)
Kent Jones (School of Science and Technology)

Abstract

In order to ensure greater consistency in grading within the Reading program at the School of Science and Technology, an online test generator named Automated Test Maker (ATM) was developed in 2016. The ATM was first developed to allow teachers to easily generate vocabulary tests of similar content and level, but with questions that are randomly chosen each time, thus, easily creating unique tests for all Reading instructors. The ATM further expanded to include a reading comprehension section, which generates final exams for the first-year Reading course. A Reading II component for final exams was recently added, but with a greater variety of tests based on the reading passages from the course textbook. Therefore, the ATM can generate different final exams for all classes using different combinations of tests based upon the studied readings. This helps to ensure greater consistency of assessment across all the Reading classes, whilst helping to circumvent the sharing of test details between classes. In this paper, the authors will report upon the following: 1) an overview of the ATM and the Reading II course, 2) the development of test items, 3) the results and analyses of the final exams, and 4) implications for the future.

1. Introduction

Quality assurance has always been demanded of higher education institutions, and with the increase in enrollment of international students, and the universalization of education, this demand has become significantly more widespread in recent times. Thus, the unification of curriculums at our institution, like many others, is encouraged and considered an important component in delivering quality education to students at all times.

Therefore, in order to meet the diverse needs of our study body, English proficiency-based

designation of classes within the School of Science and Technology at Kwansei Gakuin University was introduced in April, 2017. This applies to all three compulsory English classes that the students take during their first two years of study, namely, Communication I and II, Reading I and II, and Writing I and II. All students are allocated dependent on their GTEC scores, which they sit as a placement test before entering the university. A majority of students are then placed accordingly in either advanced or non–advanced classes. Additionally, in order to assist students who require it, remedial introductory English classes were also introduced university–wide based on GTEC scores at the same time that the new English proficiency–based class designation system was implemented. These classes are compulsory, but not conducted within our School.

Within the compulsory English classes for Science and Technology students, there are 27 classes taught by a variety of instructors. Each instructor receives course guidelines outlining learning goals and objectives, as well as the assessment criteria or minimum requirements for each program when preparing their curriculums. However, within those course guidelines each instructor is able to implement assessments of their choosing. While there are certain advantages with this current method, which allows teachers to approach their students in the way they prefer, and they can test exactly what has been taught using the specific means they choose, it can be difficult to ensure the unity of: 1. task difficulty, 2. time required for tasks, and 3. the evaluation criteria of assessments. Additionally, and very importantly, the students' GPAs have a great impact in their fourth year when they are allocated to their laboratories in the School of Science and Technology. Therefore, a level of consistency in assessment is expected by the School.

As an initial step toward greater consistency in student grading and assessment within the Reading programs, the Automated Test Maker（ATM）was developed in 2016. This was created in order to generate vocabulary tests that are unified but varied in content for an important component of the Reading program where students must obtain an average of 60 % or more over three TOEIC vocabulary tests to pass the class. Since then, about 2,000 TOEIC vocabulary questions have been created and added to improve the diversity and possible test combinations. The method for giving these vocabulary tests is now completely unified as it enables teachers to generate similar types of vocabulary tests, but with questions that are randomly chosen each time by the ATM, ensuring unique tests for each class. This system is utilized by all instructors to conduct the tests in class. The ATM is now also equipped with the computer adaptive testing function which is being implemented in certain classes before being fully rolled out. This function allows students to take vocabulary tests best suited to their level, as the questions adapt to their answers.

Furthermore, a reading comprehension section was added in 2017 to generate final exams for the Reading I course. This test bank is currently being expanded to include banks for both Reading I and Reading II courses. In the spring semester of 2019, the majority of the Reading II teachers conducted a final exam with tests generated by the ATM for the first time. In this paper, the authors will report the following: 1）an overview of the ATM and the Reading II course, 2）

the development of test items, 3) the results and analyses of the final exams, and 4) implications for the future.

2. Background

2. 1 Overview of the ATM

The ATM is a web–based program that allows users to automatically generate tests employing test items stored in the database. The test items are composed of multiple-choice TOEIC vocabulary questions and the reading comprehension test bank. To use the ATM, teachers need to input an ID and password to log–in to the website. In order to create a reading test, teachers first select "Reading Comprehension" (see Figure 1). Teachers then see the screen needed to output a reading test (see Figure 2). To do this, teachers first select multiple reading tests, then click "Add". The name of the test materials that have been selected appear listed in the right-side box. Finally, by clicking "PDF を生成," one final reading exam made up of multiple test materials based upon reading passages in the textbook is automatically generated, including blank answer sheets for students and answer keys for teachers.

2. 2 Reading Course / Final Exam

The main purpose of the Reading II course is to improve English reading abilities, with a focus on improving specific reading skills, as well as build vocabulary knowledge, using the textbook titled Core Nonfiction Reading 3 (Robinson & Alexandar, 2015) in all classes. To do this, a unified syllabus is implemented by all nine instructors who teach the 27 Reading II classes. The grading criteria is also unified across all classes, with the largest component amongst the assessments being one final exam, which accounts for 40 % of the final grade. This final exam is given to assess students' understanding of the textbook and improvement in specific reading skills. It is a paper–based test taking about 60 minutes for students to complete without a dictionary or translation notes. However, until spring 2019 it was each instructor's responsibility to create and conduct the tests in accordance to the prescribed conditions. This also included recommendations about not giving exactly the same exam to all classes if teaching more than one Reading II class in order to reduce problems that may arise with students sharing details about the tests with others. Though avoiding this problem is clearly important, most Reading II instructors teach more than one class, thus creating time management difficulties for teachers who need to create a variety of reading tests using the same pool of reading passages from the textbook or sourcing their own, whilst still teaching and assessing other classes along with carrying out non-teaching duties and responsibilities at work.

Therefore, in order to ensure greater consistency in measuring students' understanding of the materials learned and specific reading skills taught in class, the Reading II coordinators and the developer of the ATM, three of the named authors, decided to create a test bank to use with the ATM. This decision was made to ensure that all instructors would be able to easily generate

Figure 1　Log in Screen

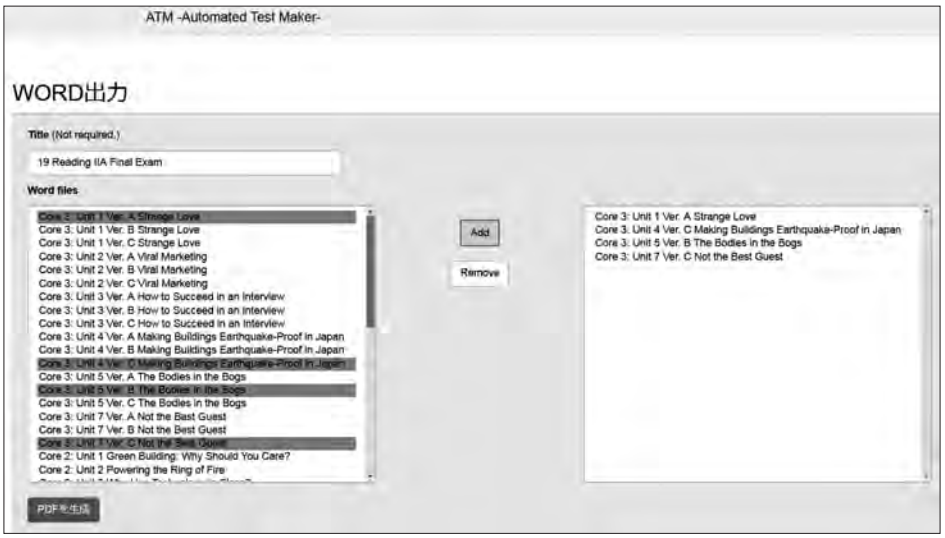


Figure 2　Reading Comprehension Test Output Screen

tests that are of similar levels and types, but are different for all classes. Thus, creating greater consistency in assessment, whilst also helping to reduce the possibility of shared details about the tests.

3. Developing a Test Bank

3. 1　Test Format and Question Types

After the decision to make a test bank was made the first important consideration was what exactly to test in the final exams and how. Though Ushiro (2012) states to precisely measure students' reading ability, a test using a reading passage that students are unfamiliar with should

be employed, he also asserts that utilizing passages already studied and questions that directly test skills practiced in class can possibly have an impact on student motivation. For example, if passages students have never read are on tests, and/or there are no questions that correspond to skills they have practiced in class, they may wonder why their exams and what they have studied do not correlate, therefore, possibly affecting their willingness to participate in the class afterwards. Koizumi, In'nami, and Fukazawa (2017) further confirm that it is preferable to use materials learned in class to check students' understanding on the content of the class, and from the perspective of students' motivation, this is favorable. Therefore, in order to maintain as much motivation as possible in a compulsory English Reading course for a student body majoring in science and technology, it was decided to utilize the reading passages studied in the class for the creation of the test bank at this initial stage.

After deciding to employ the reading passages from the textbook, three of the authors created a number of pilot tests referencing the already created reading comprehension exam component of the Reading I course. These first versions of the tests had various types of questions such as multiple choice, choosing true/false/not given, cloze with and without choices, short answers, and so forth. Careful attention was paid to each question type to ensure that it assessed the objectives stated on the unified syllabus, and that each question reflected question types and skills in the textbook, Core Nonfiction Reading 3. Therefore, though they were initially considered, questions requiring short written answers and essays were eliminated. Finally, it was decided to use two question types that assessed skills studied in the textbook chapters, namely multiple–choice questions and fill–in–the–blanks that require students to find words or phrases within the reading passage to finish a set of information.

Not all textbook units were studied during the spring semester, therefore at this stage tests were created only for those units covered. In advanced classes, seven units of the textbook were required to be studied during the spring, while six units or more were assigned for non–advanced classes, with each passage being around 350 words long. It was decided to maintain the 60 minutes for testing that had previously been prescribed. Consequently, it was not possible to test every unit studied, however it was determined that it would be feasible to test more than half of the units. Given that, it seemed reasonable that the number of questions for each unit fall between 10 to 12, with each test finally ending with 12 questions. Multiple tests for each unit studied during the spring semester were then created with 12 questions each utilizing the question types mentioned above.

3. 2 Creating Test Items

Important points that were considered when creating tests for each unit are summarized below. All tests should:

1) assess the students' understanding of the reading passages as well as six specific reading skills studied in class: categorizing, cause and effect, fact and opinion, problem and

solution, terms and descriptions, and sequencing,

2) make sure to use question types that students are familiar with and are studied in the textbook; multiple choice or fill–in–the–blanks,

3) have 12 test items per test with more multiple–choice questions than fill–in–the–blanks, and

4) have three different versions per one unit.

With this in mind the two Reading II course coordinators, who are also two of the authors, created three different tests for each of the seven units prescribed to be studied during the spring semester, making a total of 21 tests. This was done to ensure that random combinations of tests from different units could be utilized to create unique final exams whenever needed.

A typical multiple–choice question created during this stage is as follows:

Question 10: What can you infer from the underlined sentence **10**?

a. Animals are fortunate because they do not need to be single like some humans.

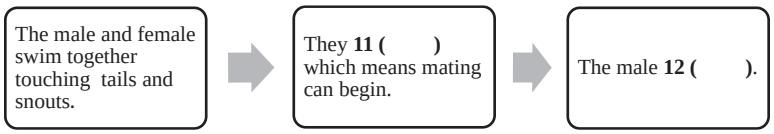
b. Animals are comfortable because they have the ability to attract mates, unlike some humans.

c. Humans are fortunate because they do not need to perform unusual rituals to find mates like some animals.

d. Humans are comfortable because they do not need to be single to perform mating rituals.

Typical fill–in–the–blank questions created during the initial phase described above are as follows:

Questions 11 and 12: Complete the sequence for the seahorses' mating ritual. For each answer, choose a word or phrase from the reading passage. Use NO MORE THAN FIVE WORDS from the reading passage.



After all the tests were created, the other authors, who are all test bank project members, proofread all the 21 tests as well as created the answer keys. During this stage some minor mistakes like spelling errors were found and fixed, and items that had more than one answer were modified. Important issues with some of the fill–in–the–blanks questions were also discovered during the proofreading phase. As the examples above show, students were instructed to "use" no more than five words from the reading passage to complete the information. However, this was deemed not clear enough, causing one of the authors to detect more than 15 possible answers for one question. To prevent this problem the instructions were altered to include, "DO NOT change

Table 1 Unit 3 Test Item Distribution

Ver.	Test item number											
A	0301	0302	0303	0304	0305	0306	0307	0308	0309	0310	0311	0312
B	0313	0314	0315	0305	0316	0317	0307	0318	0309	0319	0320	0321
C	0313	0302	0316	0315	0303	0306	0322	0323	0319	0311	0321	0324

Note. Four-digit numbers indicate a test item. The gray highlight shows fill-in-the-blanks questions while others are multiple choice.

the word/phrase form or order."

3. 3 Final Test Items

After several proofreading sessions, test items were finalized. Each textbook unit had three different tests, versions A, B, and C, with each version containing a different combination of 12 questions. Therefore, each unit had 36 questions, and whilst some questions were unique to each test, some questions were utilized in multiple tests. Table 1 shows the distribution of test items across versions A, B, and C for one unit. As can be seen with the underlined and double underlined numbers several questions are shared between versions. Namely, version B shares the underlined test items 0313, 0315, and 0316 with version C, while version A shares the double underlined test items 0307 and 0309 with version B. However, as can also be seen, all three versions are not identical and contain questions unique to each. Therefore, from the 252 questions in total across the 21 tests, 177 of these were uniquely created for the test bank with some used multiple times due to this overlapping process. It should also be noted that the authors did not intend to make one test version for each unit more difficult than the other two, but rather tried to keep all the tests at the same level of difficulty within all units.

As explained earlier, the final exams were intended to test more than half of the required units for each class, that is, four units for non–advanced and five units for advanced classes, thus one final exam consisted of one test version from four or five units, containing 48 or 60 test items in total. In order to do this, all of the 21 tests were uploaded to the ATM, making it possible to easily generate different final exams for all of the Reading II classes.

4. Conducting Final Exams

Using the test bank to create a final exam was not required by all of the Reading II teachers during the initial phase being described. However, using the test bank was offered to all Reading II instructors, with two not involved in the project opting to do so, meaning that the tests were utilized to conduct a final exam in 23 Reading II classes out of 27. During this initial phase of testing, it was important that all of the 21 tests be utilized. This would allow the authors to check test item difficulties and identify any unforeseen issues with the tests. This of course would contribute to the development of the test bank for the following semester and into the future.

Therefore, instead of each teacher accessing the ATM to generate a final exam, final exams were created by one of the authors in order that all 21 tests would be fairly and equally distributed among all 23 classes, but none would be identical.

It should be noted that each created test had two to six fill–in–the–blank questions, thus it was carefully arranged so that the number of fill–in–the–blank questions ranged from a total of 22 to 24 out of 60 items in advanced classes, and 15 to 18 out of 48 questions in non–advanced classes. Furthermore, it was made sure that each final exam included more than three different kinds of the six specific reading skills studied in class. For future development in setting item difficult parameters, Unit 2 tests were utilized in all 23 final exams.

All the final exams were conducted either on the last day of the course or the week before, depending on an instructor's preference. A total of 526 students, 197 in advanced classes and 329 in non–advanced classes, took an exam utilizing the test bank created on the ATM. As mentioned earlier, 60 minutes was allocated for all tests, and students were not allowed to use any resources such as a dictionary or translation notes. All the final exams and answer sheets, whether used or unused, were counted, collected, and returned by the instructors to the Reading II coordinators.

Once all the materials were returned to the coordinators, the answer sheets were scanned for multiple–choice questions, and then the fill–in–the–blanks sections were marked by five of the current authors. As for marking the fill–in–the–blank questions, certain guidelines were decided upon by the attending project members before grading proceeded. Importantly, though every effort was made to make instructions very clear, it was decided to allow partial points (0.5) in certain cases that showed reading comprehension and skill, but did not adhere exactly to the instructions. Partial points were given in the following cases despite the instructions asking students to use the exact word(s) in order from the reading passage within the word limit given: 1. missing an article, 2. missing –s for plural or verb, 3. wrong form, 4. close meaning but not precisely correct, and 5. (an) extra word/s from the reading, but within the limit for the correct answer. All versions of one unit were marked by the same project member for consistency, meaning that multiple members checked different parts of each final exam. Furthermore, one of the authors double checked all of the final exam answer sheets later when calculating the results of the multiple–choice questions and fill–in–the–blanks. The final checker then informed each Reading II teacher about the results of the final exams.

5. Results

Tables 2 and 3 show the descriptive statistics for the results from the final exams conducted in nine advanced classes, and 14 non–advanced classes. A web–based assessment tool (Mizumoto, n.d.) was employed to calculate the results in Tables 2 and 3. It should be noted that the class numbers used in this article are not the actual class numbers used in the English program. These numbers have been randomly assigned for this article to ensure the anonymity of all participants.

Table 4 shows the means of each test version included in the final exams for each class. For

Table 2 Results of the Final Exams in Nine Advanced Classes

Class	*n*	*M*	*SD*	*Median*	*Min*	*Max*	*Cronbach's α*
1	22	44.02	9.00	46.75	22.50	58.00	0.89
2	23	42.02	7.62	41.50	28.00	54.00	0.86
3	18	39.03	7.99	38.75	22.00	51.50	0.85
4	21	42.93	5.77	43.50	31.00	52.00	0.75
5	25	36.52	10.02	33.50	21.00	56.00	0.91
6	22	40.89	7.67	40.25	26.50	53.50	0.86
7	23	44.30	5.33	46.00	34.00	50.00	0.71
8	18	45.03	7.17	45.00	33.50	58.50	0.84
9	25	42.04	6.30	42.50	27.50	50.50	0.80

Note. For advanced classes, the maximum point was 60.

Table 3 Results of the Final Exams in 14 Non-advanced Classes

Class	*n*	*M*	*SD*	*Median*	*Min*	*Max*	*Cronbach's α*
10	19	25.16	7.08	23.50	14.00	38.00	0.85
11	24	29.67	6.22	30.00	18.50	39.00	0.81
12	22	33.52	6.24	34.00	22.00	45.00	0.81
13	23	28.04	7.11	28.00	15.00	40.00	0.86
14	25	29.67	6.22	30.00	18.50	39.00	0.81
15	24	31.79	6.52	31.00	23.00	44.50	0.83
16	25	28.04	8.06	28.00	14.00	41.00	0.86
17	23	26.59	6.11	27.50	11.50	34.50	0.78
18	20	34.70	8.60	39.25	20.50	46.00	0.90
19	24	25.29	6.53	23.25	14.00	42.00	0.80
20	29	26.47	8.64	24.50	11.00	45.50	0.89
21	18	29.08	7.47	29.75	40.50	16.50	0.86
22	26	30.60	6.68	30.75	17.50	41.00	0.81
23	27	31.61	5.38	31.25	20.00	40.50	0.75

Note. For non-advanced classes, the maximum point was 48.

example, Class 1 (as randomly named for this paper) took a final exam composed of Units 1, 2, 3, 6, and 7 consisting of versions C, A, C, C, and A respectively. The maximum score for each test version was 12. Therefore, in this class the highest mean was for Unit 3 (version C), whilst the lowest was Unit 7 (version A).

Next, of the 177 original questions created for the test bank, there were two question types; 74 fill-in-the-blanks and 103 multiple-choice questions. Figure 3 shows the correct answer rate for each type. Question numbers are arranged by the correct answer rate from the lowest to the highest in order to see the results of each test item clearly. As can be seen, both of the correct answer rates are distributed similarly with the multiple-choice questions resulting in slightly higher correct rates than the fill-in-the blanks. Furthermore, the results indicate that some

Table 4　Results of Each Unit and Version

Class	1A	1B	1C	2A	2B	2C	3A	3B	3C	4A	4B	4C	5A	5B	5C	6A	6B	6C	7A	7B	7C
1			9.0	9.0					9.6									8.3	8.0		
2	9.1			9.0			8.0				6.1									9.8	
3		8.4				6.7		9.3									7.3			7.4	
4				9.1			8.5				5.8		8.7							10.9	
5			8.3		6.8				8.8	5.6									7.0		
6					7.8				9.3			6.4		9.0					8.4		
7		9.9				8.0		9.5				7.6									9.3
8	9.3				8.8		8.5									8.8					9.7
9						8.5		9.7		7.2					8.5						8.2
10	8.5					5.3						5.2	6.2								
11						6.8	7.6								7.8		7.5				
12			8.0	9.2										8.9				7.4			
13		9.4			7.0					5.2					6.5						
14	9.5					6.8									7.8		6.3				
15		9.9			7.4										8.0	6.5					
16				6.8					8.1								6.6		6.2		
17			7.2	7.6							4.5			7.3							
18			8.8	8.8										9.0				8.1			
19				6.3				7.9					5.8			5.4					
20		8.5			5.6								6.5			5.8					
21					5.8				8.0					8.8				6.5			
22				7.6				8.2								7.1					
23	9.0					6.7							8.7				7.2				
Mean	9.1	9.2	8.3	8.1	6.9	7.0	8.1	8.9	8.8	6.0	5.5	6.4	7.2	8.6	7.7	6.6	6.9	7.6	7.4	9.5	8.6
Ss #	112	117	112	206	160	161	86	116	112	73	67	64	120	105	121	121	119	82	94	62	92

Note. The number in the top column shows unit number and the letter indicates its version. The maximum point is 12 for each test. The number indicates the mean. Gray indicates the highest score of a unit test within each class' final exam whilst black indicates the lowest. Mean indicates the mean of all of the students who took that version of the test within the final exam. Ss # means the number of students who actually took that portion.

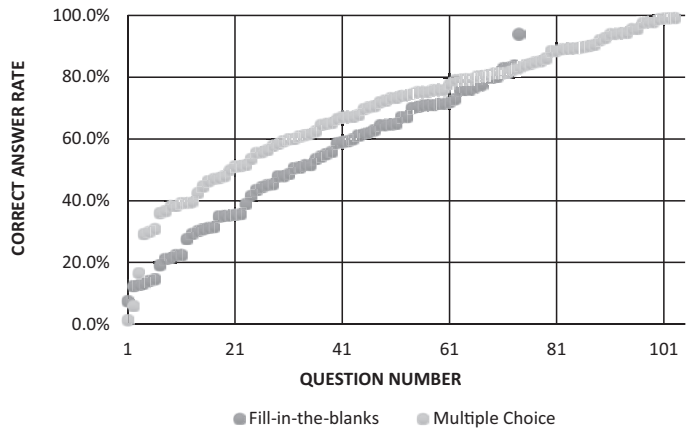


Figure 3 The Correct Answer Rates of Fill-in-the-blanks and Multiple-choice Questions.

questions clearly received more correct answers than others, suggesting a variety in the level of difficulties for the test items. Thus, the created exams had a combination of easier to answer and more difficult to answer questions. However, there are a few outliers that potentially need to be addressed before those items are once again utilized.

Furthermore, these 177 questions can be divided into the following two areas that the tests were intended to assess; general comprehension of the reading passages studied in class and the six specific reading skills mentioned earlier. The former questions can further be arranged into seven classifications: fact/negative fact (15 items), inference (3 items), understanding the flow of the passage (15 items), key vocabulary (15 items), pronouns (6 items), details (47 items), and main idea (15 items). And, as can be seen in Figure 4, where question numbers are arranged by these types by correct answer rates from the lowest to the highest, regardless of which kinds of questions they were, they again are mostly evenly distributed from easy to answer to difficult to

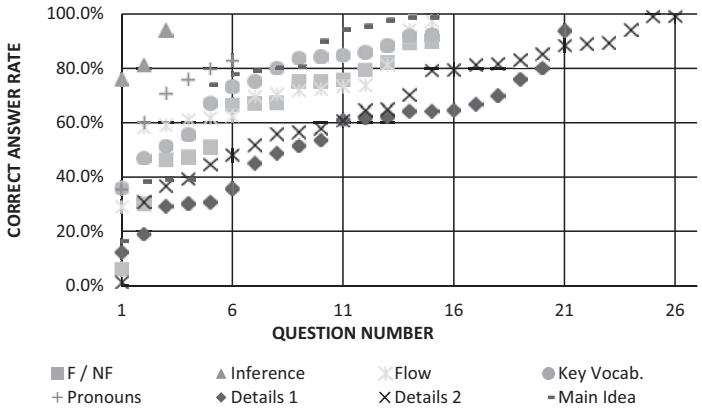


Figure 4 The Correct Answer Rates of the Questions to Assess Students' Understanding of the Reading Passages. F/NF = Fact/Negative Fact. Details 1 = fill-in-the-blanks to ask about the details. Details 2 = multiple-choice questions to ask about the details. The other kinds include both question types.

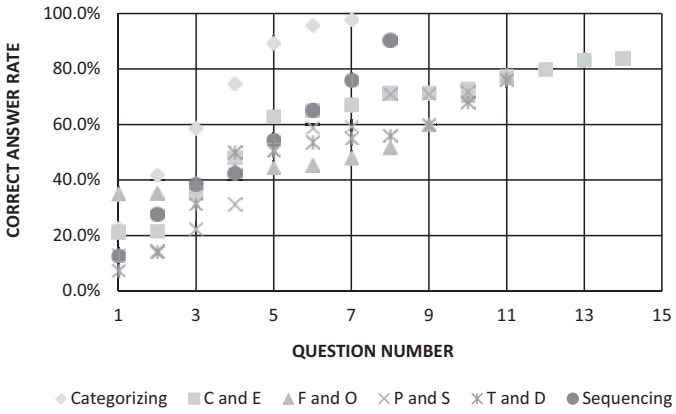


Figure 5　The Correct Answer Rates of the Questions for the Six Specific Reading Skills. C and E ＝ cause and effect. F and O ＝ fact and opinion. P and S ＝ problem and solution. T and D ＝ terms and descriptions.

Table 5　Top and Bottom Five Questions

Unit	Version	FiB/MC	Item Type	Rate
2	C	MC	Details	1.2%
4	B	MC	Fact	6.0%
5	C	FiB	T and D	7.4%
4	A	FiB	Details	12.3%
7	A	FiB	Sequencing	12.8%
5	A	MC	Details	97.8%
5	B	MC	Details	98.8%
3	A/C	MC	Main Idea	99.0%
3	A	MC	Main Idea	99.0%
1	A/B	MC	Categorizing	99.2%

Note. FiB = fill-in-the-blanks. MC = multiple-choice questions. T and D = for Terms and Descriptions.

answer. The other area of questions tested the six specific reading skills that the students studied in the textbook during the semester. Figure 5 shows the correct answer rates of each reading skills question arranged from the lowest to highest, again showing a steady distribution of difficulty, however, overall the questions for categorizing achieved higher results than the other skills. Finally, the top 5 least correctly and most correctly answered questions are listed in Table 5.

6. Discussion

Because each exam employed for the 23 final exams was different from each other and taken by different groups of students, they cannot be exactly compared nor generalized. However, the mean of each exam ranged from 36.52 to 45.03 in advanced classes and from 25.16 to 34.70 in non–advanced classes (see Tables 2 and 3). Overall, the students in advanced classes performed

better than the ones in non–advanced classes even though their exam had 12 more questions within the same amount of time.

As Table 4 shows, all versions of the Unit 4 tests were the most difficult, with the tests resulting in the lowest scores when compared to the other tests. In the future, Unit 4 may need to be altered to adjust its level of difficulty when compared to other units, or noted as a challenging level needing to be combined with other units that resulted with higher levels of accuracy such as Units 1 and 3. Similarly, this may also indicate that these units with greater levels of accurate results need to be adjusted or should be combined with more difficult tests.

Figures 3, 4, and 5 all confirm that when considering all question types within the fill–in–the–blanks and multiple–choice question types, no matter which kinds of question items they were, the correct answer rates were spread similarly, indicating that the all tests included easy, moderate, and difficult items. Furthermore, despite earlier concerns by project members that multiple–choice questions would be found much easier than fill–in–the–blanks, it became apparent that this test bank included a variety of difficulties for students in both question types.

Regarding test questions that were meant to check students' general understanding of the reading passages, there were only three items labeled as inferences and six items as pronouns. When compared with the rates of the other question kinds this disparity will need to be addressed in future test bank edits and editions. Furthermore, as can be seen in Figure 4, the number of correct answers for questions relating to the main idea of the text or paragraph, and vocabulary questions, were higher than for other kinds of multiple–choice questions. However, it is worthy to note again, that within the same question type, the distribution between most and least correctly answered was mostly widely, but evenly spread.

Next, it was expected that results would indicate a specific reading skill that students were obviously more adept at than others. As Figure 5 shows, questions with categorizing received higher results than the other skills, but there was not a skill that students were clearly more capable of or not.

Finally, in Table 5, both the top and bottom five most correctly answered questions are listed, resulting unexpectedly with the bottom two being multiple–choice questions. This result has compelled us to revisit the questions and consider their difficulty before being employed again in future exams. Additionally, as can be seen, the bottom items come from different units and are all different question kinds. On the other hand, the top five items are all multiple choice, which was expected, and similar types of questions that asked for some kind of general understanding of the passage. Thus, indicating students' reading strengths currently lie in this area.

Again, these test items were answered by different students from different majors, so it is difficult to conclude anything universal from the results of these final exams during this initial run of the reading test bank within our School. However, it does assist the authors in understanding if the students reached their learning goals within the Reading II course and what they have accomplished not just in the class(es) one teacher teaches, but as a whole.

7. Conclusion

This paper was written in order to introduce the ATM, and to share how a test bank for one Reading course was initially developed. As of March, 2019, when this project first commenced, there was no previously published research that the authors are aware of about the development and implementation of a test bank using an in–house developed web–based program to generate different but similar types of tests automatically. While it was successfully completed and we learned there are advantages to this method as described, there are several issues to be addressed.

One aspect of the tests that needs to be re–considered regards the instructions for the fill–in–the–blank questions. Though instructions were clearly stated and we assumed students would be familiar with them as they encounter these kinds of instructions during their tests taken at high school or entrance exams, some students were still able to find unanticipated correct answers or answers that fit but did not follow the instructions of not changing the form or order of the words from the passage. Despite these instructions being added to avoid an over–abundance of possible correct answers, students on occasion demonstrated the necessary reading comprehension with responses that answered the question, thus receiving partial grades as described earlier, but were technically incorrect in regards to the instructions. Therefore, these directions will need to be carefully reviewed. However, importantly, it should also be noted that in the future fill–in–the–blanks will be manually marked by the instructor of each class just as they were by the project members in this study. Consequently, these kinds of unanticipated answers will be considered individually by each instructor and marked accordingly. Furthermore, guidelines for grading these kinds of answers, much like those agreed upon by the project members at marking during the phase currently being reported on, will be formulated to help ensure a level of uniformity and consistency with the grading across the Reading program.

Finally, as happens with any initial phase of a project, certain elements that require editing or adjusting were also discovered. Therefore, though it was not necessary for all of the tests, after grading the tests were returned to the proofreaders, now with further insight after implementing and marking tests, for review and to address any potentiall problematic areas.

As this was an initial phase to develop a test bank for the Reading II course, it was not expected that a majority of teachers would voluntarily use the test bank when it was launched. However, despite it not being a requirement, they did so. Indicating that most instructors within the Reading II course desire greater consistency in grading and assessment. Also, it should be noted that one instructor who did not conduct a final exam using the test bank had a great interest in joining the project, however, had wished to use reading passages that the students were unfamiliar with for the final exams. In the creation of exams for these classes the Reading II coordinators were consulted in order to discuss the test format, amount, and difficulty level, to ensure they were as consistent as possible with those from the test bank. In the future it is hoped that tests based on new reading passages that still test skills practiced in the textbook, perhaps

with similar topics to those studied in class, will be added to the test bank to address the needs of teachers who prefer this method, whilst also increasing the possibilities available with the test bank on the ATM. Combining tests with reading passages from the textbook and tests with new/similar articles will not only allow us to assess students' reading ability more widely, but also still maintain the students' motivation to study diligently in class. Eventually, we envisage all the reading teachers utilizing the test bank to ensure greater consistency in assessment across the Reading programs.

References

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambrige: Cambridge University Press. 石川祥一・西田正・斉田智里 (2011).『テスティングと評価』大修館書店.

笠原究・佐藤臨太郎 (2017).『英語テスト作成入門』金星堂.

小泉利恵・印南洋・深澤真 (2017).『実例でわかる英語テスト作成ガイド』大修館書店.

Mizumoto, A. (n.d.). *langtest.jp* Retrieved from http://lantest.jp.

根岸雅史 (2017).『テストが導く英語教育改革』三省堂.

Nuttall, C. (2005). *Teaching Reading Skills in a Foreign Language*. Macmillan.

住政二郎・工藤多恵・乗次章子・山脇野枝 (2019).「自動テスト生成システム (ATM) の開発と実践への応用」『関西学院大学高等教育研究』9. 19–26.

卯城祐司 (2012).『英語リーディングテストの考え方と作り方』研究社.