Effectiveness of an Online Language Learning Platform in China

Ryan Baker, University of Pennsylvania

Feng Wang, Learnta, Inc.

Zhenjun Ma, Learnta, Inc.

Wei Ma, Renmin University of China

Shiyue Zheng, Teachers College Columbia University

Abstract

In this paper we evaluate the effectiveness of an adaptive online learning platform, designed to support Chinese students in learning the English language. The adaptive platform is studied in three studies, where the experimental platform is compared to an alternate, non-adaptive platform, with random assignment to conditions (the adaptive learning platform versus a control non-adaptive online platform covering the same content). The three studies had 39 students, 91 students, and 90 students, all in the eighth grade but in different schools in China. The two conditions were compared in terms of gains on pre- and post-tests relevant to the local schools, as well as on a survey measuring student attitudes towards the technology. The adaptive platform is found to lead to better and faster learning than the non-adaptive platform, as well as more positive responses to survey items.

Studying the Effectiveness of an Online Language Learning Platform in China

Developing competence in languages other than one's native language is increasingly a goal for students worldwide (Kasteen, 2014), as the global economy and competition become increasingly worldwide phenomena. Online instruction has become a prominent part of second language learning, often referred to as computer-assisted language learning (Beatty, 2013). Popular systems such as Duolingo (von Ahn, 2013), other systems. While online second language learning started in Western countries, perhaps most notably with the PLATO system (Marty, 1981), this trend has grown worldwide, with online language learning becoming prominent in Brazil (Abreu-Ellis et al., 2013), China (Macaro et al., 2012), and many other countries.

Online language learning has several advantages over other forms of instruction, including exposure to native pronunciation, which can be difficult to find in many countries with live instructors (Levis et al., 2016). In addition, across domains, online learning offers the promise of adaptation to individual learner differences, leading to better and more efficient learning (Kulik & Fletcher, 2016). China in specific has seen a number of online platforms appearing which propose to support second language learning, including YuanTiKu (http://www.yuantiku.com), MoFangGe (http://www.mofangge.com), New Oriental Online (http://www.koolearn.com), eXueDa (http://www.exueda.com), XueErSi (http://www.xueersi.com), and Classba (http://www.classba.com).

However, as with many of the online and adaptive learning products that have sprung up in the American market over the last decade, it is unclear how effective many of these products are. No external stamp of quality is required to proclaim one's product effective or adaptive (cf. Kroeze et al., 2015). In the United States, this has led to greater interest in experimental studies,

spurred by the role played by the What Works Clearinghouse in K-12 education (http://ies.ed.gov/ncee/wwc/), leading to large-scale randomized controlled experiments evaluating the effectiveness of online learning systems (e.g. Pane et al., 2014; Roschelle et al., 2016).

However, efforts to establish effectiveness are still more nascent in the Chinese educational market. There have been single-condition studies showing positive motivation among students using computer-assisted language learning systems in China (Sun, Zhang, & Dong, 2003; Zhao, 2015), but without comparison to a control condition, it is difficult to know whether these solutions are genuinely better than more traditional approaches. There have been studies showing that teacher-supported online learning platforms (i.e. with discussion forums and participation with classmates) can be more effective at promoting learning in China than traditional teaching methods (Wang, Liu, & Fu, 2015; Li, 2016), but this is rather different from the adaptive learning platforms being increasingly used across China.

This is a problem in two key ways. First of all, many students may be using products that are less educationally effective. Second, it is not clear that many of the adaptive features that have been tested and validated within Western populations will be equally effective in Chinese populations. Many adaptive features that were thought to be highly valuable at first have not been shown to actually lead to differences in student outcomes in the United States (VanLehn, 2006). It is known that adaptive learning environments are often used in different ways by students in different cultures and countries (e.g. Ogan et al., 2012; Rodrigo et al., 2013). Therefore, it is worth asking whether forms of adaptivity that have been shown to be effective in the United States, such as using knowledge spaces to select items (cf. Doignon & Falmagne, 2012), will be equally effective for Chinese students.

We research these questions in the context of three studies investigating whether the Adaptive Learning System (ALS), developed by Classba.com, will be more effective, less effective, or equally effective as an alternative popular online learning platform which teaches the same content. Three distinct studies are presented, with comparable design goals but different implementational designs (in line with the practices at the three educational settings they are conducted in). There is increasing awareness of the value of integrating across multiple studies investigating the effectiveness of the same intervention (see in this type of system, for instance, work to study the impact of the ABRACADABRA reading system across Canada, Australia, and Kenya -- Abrami et al., 2015). In line with this practice, we use research synthesis methods to integrate results across the three studies, towards answering the research question: Does the Adaptive Learning System lead to better and more efficient learning than a popular non-adaptive platform? We also investigate differences in student attitudes between the two conditions.

Systems Compared

In this paper, we compare the effectiveness of the Adaptive Learning System (ALS) to a control condition, New Oriental Online Learning (also referred to as koolearn or EDU). Students were randomly assigned to use either system. In this section, we describe the two platforms.

We compared the two systems within a lesson on passive voice in the English language. This topic was selected based on its importance for successful academic and professional communication, and its known difficulty for students in China (Chuang & Nesi, 2006). Within the Chinese market, there are several online learning products teaching passive voice to elementary and middle school students, including the platforms discussed in the introduction. New Oriental was selected as the control condition, as New Oriental's passive voice course is the

closest fit with the Adaptive Learning System's passive voice course in terms of content style and intensity.

Within New Oriental, students completed the "2 hours nailing down passive voice" curriculum. Learning content included four declarative lecture videos on passive voice; these four videos had a total duration of 86.5 minutes. In addition, students viewed one 17-minute video of worked examples, where the instructor provides a set of worked examples on how to correctly complete questions. The system also gave the student 50 assessment items. The student could attempt each item, and after the student answered, the system gave the student the correct answer and provided a textual explanation (whether or not the student was correct). Students could access the content in any order but most students followed the prescribed order of watching the four declarative lectures and then watching the worked examples, prior to completing the assessments. The amount of curriculum content provided was controlled for, and students completed it at their own pace. Students were not required to watch the entire videos, in line with the usual usage practices for this system, and the learning time taken by students in this study ranged from 50 minutes to 180 minutes.

Figure 1. The New Oriental Online (koolearn) platform.

The learning content provided by the ALS involved material corresponding to the same topics within passive voice grammar as the material studied within New Oriental. However, the content itself was developed separately. The content in ALS consisted of 24 passive voice declarative lecture videos which had a total duration of 89.6 minutes, plus a set of practice items.

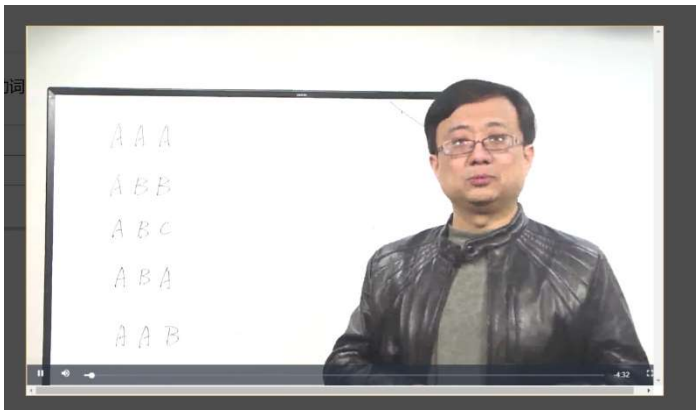

Figure 2. The Adaptive Learning System.

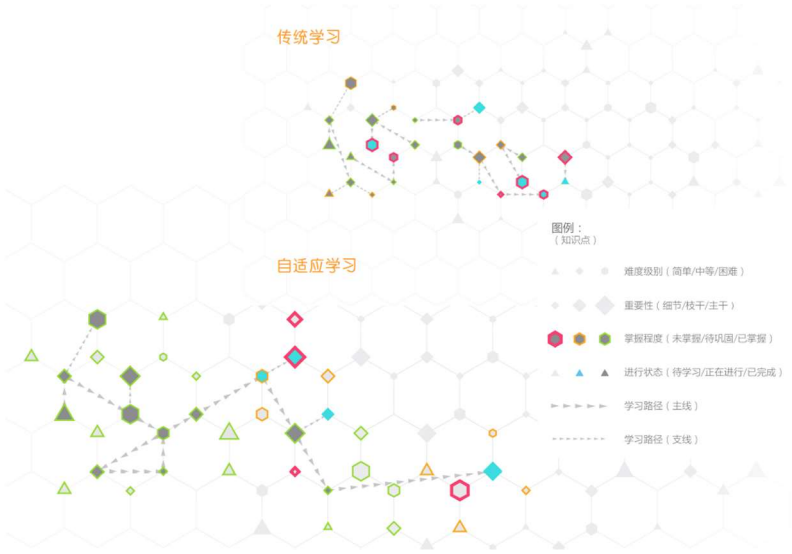

Figure 3. The Adaptive Learning Path, within the Adaptive Learning System. This shows the student's progress through the curriculum.

However, the way that content was provided was quite different in ALS than in New Oriental. ALS starts with a diagnostic test of 25 items that develops an initial estimate of student knowledge of 23 concepts pertaining to passive voice. At the end of the diagnostic test, a list of mastered concepts and a list of unmastered concepts are generated. These lists are presented to students for their review, along with statistics on their performance (such as percentage of items correct), and a list of answered items and explanations of the correct answers to each item.

After reviewing these explanations, the student enters the second phase of their use of the system: focused learning and practice. In focused learning and practice, a student works on the unmastered concepts one by one. ALS determines the next concept that the student will be asked to learn among the list of unmastered concepts (learning path recommendation). For each unmastered concept to learn, a student is offered a video (each video is focused on a single concept), a text, two practice items with solutions and explanations available immediately after

the student answers, and up to five assessment items with solutions and explanations available only after the completion of the assessment. In the current version of ALS, neither practice items or assessment items have hints. ALS determines what a student will get next and the student only need to click the "Next" button. Item Response Theory is used to determine the difficulty level of practice and assessment items.

After the assessment of a concept, the concept is classified as either mastered or not mastered, and the student's knowledge state is updated. ALS will then recommend the next unmastered concept for the student to work on. If a student fails to master a concept in the first attempt of focused learning and practicing, the student would get a second chance on this still unmastered concept after working on at least two other concepts. In that second chance, a student might be given the same video to watch again but would receive new items for practice and assessment.

Students were not required to watch entire videos, and the learning time taken by students in this study ranged from 18 minutes to 180 minutes.

Experimental Procedure

Participants

This study was conducted in three schools, which were selected to represent some of the common settings of use for K-12 online learning in China. The three schools were each located in a different large city in Eastern China. 39 eighth graders participated in the study from School A, which is a private after-school institute in the center of a large Chinese city. It provides after-school programs with classes in various subjects. School B and School C are both public secondary schools located in the center of large Chinese cities. 91 eighth graders from School B

and 90 eighth graders from School C participated. All three schools include students from a range of economic backgrounds. All three schools are composed almost entirely of students who speak Mandarin as their native language.

Each participating student and their parent signed a consent form for participation in the study. Any student who did not have these consent forms did not participate in the study, and instead participated in an alternate activity. Any student was allowed to quit the study at any time. Each student's data was recorded with reference to an anonymous ID, such that it was not possible to re-identify any participant after the study.

Data was collected on each student's background information, including gender, age, whether the student intends to attend a competitive high school, whether they have previously studied English passive voice, whether they like learning English, their comfort with computer size, their parents' education level, their perception of how much their parents value their education, their satisfaction about their family's economic status, and their expectation of whether and how much they will improve on high school entrance examinations after using the learning system.

Students in School A participated in the study during the 2016 winter break (January-February). They took a ten-day course where the first seven days were their regular supplementary courses at School A, and participated in the experiment during the last three days. These students studied passive voice in their classes prior to participating in the study. For them, the passive voice topic in this experiment was review rather than learning new grammar knowledge.

Students in Schools B and C participated in the study during two class days within the same week, in March 2016. These students had encountered passive voice during their previous

studies, but their classes had not systematically covered passive voice prior to the study. As such, the passive voice topic in this experiment can be considered to be new learning rather than review.

After completing the study, all students completed a 17-item questionnaire designed by researchers in China, covering their perception of the effectiveness of the system, its usability, and their satisfaction of the system.

Test Design

Student learning was evaluated with pretests and posttests. As the three schools were in different provinces of China, the pretest and posttest exams were not the same across all schools, but were based on the high school entrance exam relevant to each school. Attempts were made to ensure comparable difficulty between the pretests and posttests, but counterbalancing was not used due to local implementational challenges. Item formats included multiple choice, fill in the blank, and sentence structure transformation. In School A, a local curriculum expert selected items from a test pool of items with tagged content strands and known difficulty levels identified by other local curricular experts, randomly selecting items within each strand to be on either the pretest or posttest. In School B, an expert teacher generated the two exams to be of equal difficulty, and then the exams were separately verified by a local curriculum expert to be of comparable difficulty. School C decided to use the same exams as School B after review by local curriculum experts. The exams used by School A had 50 items; the exams used by Schools B and C had 45 items. All exams have a maximum score of 100, scored as the percentage of items answered correctly.

Study Protocols

Moderately different study protocols were applied at each school, in line with the differing conditions in each school. In School A, students were given 30 minutes to take the pretest exam on paper, and then were given 20 minutes to become familiar with using the tablet computer and the learning system assigned to their condition. Afterwards, the students used their assigned system for 60 minutes a day for 3 days, for a total of up to 180 minutes of learning time (as discussed above, some students completed in less time than 180 minutes). Teachers and researchers did not interact with the students during use of the software except for technical problems, in order to focus this study on the effect of the systems themselves on learning. Both the control and experimental groups participated in the study within the same classrooms under the supervision of the same teachers. Students filled out their background information questionnaire on the tablet in the 5 minutes before the end of the course on the first day of the experiment. At the end of the third day of class, all students were given 30 minutes to take the posttest exam on paper. After the posttest, students answered a questionnaire about the system on the tablet.

In School B, students were given 25 minutes to take the pretest exam on a desktop computer, and then were given 20 minutes to become familiar with using the learning system assigned to their condition on the same desktop computer. Afterwards, the students used their assigned system for 50 minutes, 50 minutes, and 40 minutes across 3 days, for a total of up to 140 minutes of learning time. Teachers and researchers did not interact with the students during use of the software except for technical problems, in order to focus this study on the effect of the systems themselves on learning. Both the control and experimental groups participated in the study within the same classrooms under the supervision of the same teachers. Students filled out

their background information questionnaire on the computer in the 5 minutes before the end of the course on the first day of the experiment. At the end of the third day of class, all students were given 25 minutes to take the posttest exam on the computer. After the posttest, students answered a questionnaire about the system on the computer. Some students at this school also participated on interviews about ALS or answered open-response items about ALS after the study, which will not be discussed in this paper.

In School C, students were given 25 minutes to take the pretest exam on a desktop computer, and then were given 20 minutes to become familiar with using the learning system assigned to their condition on the same desktop computer. Afterwards, the students used their assigned system for 50 minutes each day across 2 days, for a total of up to 100 minutes of learning time. Teachers and researchers did not interact with the students during use of the software except for technical problems, in order to focus this study on the effect of the systems themselves on learning. Both the control and experimental groups participated in the study within the same classrooms under the supervision of the same teachers. Students filled out their background information questionnaire on the computer in the 5 minutes before the end of the course on the first experimental day. At the end of the third day of class, all students were given 25 minutes to take the posttest exam on the computer. After the posttest, students answered a questionnaire about the system on the computer.

Analytical Methods: Learning Gains

In the following sections, we will compare learning gains in each of the two conditions, as well as learning efficiency (learning gain per time), comparing between ALS (experimental)

and New Oriental (control). We will also analyze the results of the questionnaires administered to students, looking for differences between ALS (experimental) and New Oriental (control).

We conduct the statistical analysis of the learning gains in these three studies in two steps. First, we conduct a statistical analysis of whether there was learning in each study. Second, we aggregate across the three studies using Stouffer's Z (Mosteller & Bush, 1954) to examine the overall effect of using ALS. Stouffer's Z is chosen rather than an approach that controls for group-level variance (i.e. a multi-level model or hierarchical linear model) because the different protocols and different pre-tests and post-tests used in the three studies properly make this three separate studies (albeit with relatively similar designs) rather than a single study at three sites. While Stouffer's Z is most commonly used in classical meta-analysis, where a researcher integrates across a larger number of studies conducted by different research groups (e.g. Orwin, 1983; Lipsey & Wilson, 2001), it remains valid for the case of integrating across a small number of studies conducted by the same research group (Rosenthal & Rosnow, 1991).

For each of the studies, we conduct a Repeated Measures ANOVA. In this analysis, each student is assigned to one of the between-subjects treatment levels (experimental and control). The within-subject factor is time (pre-test and post-test), and the between-subjects factor is online learning system (ALS and New Oriental). The sources of variation are composed of between-subjects (learning system/condition) and within-subject effects. Within-subject effects includes the main effect of within-subject (time) and its interaction with the between-subjects factor (condition* time). We examined each school's test scores for violations of normality. After verifying that kurtosis and skewness statistics were within acceptable bounds, we used a repeated measures ANOVA for the individual study analyses. We used a repeated-measure

design because we are interested in the effect of the between-subjects treatment over time, that is, the learning system/condition * time interaction.

We then integrate across the three studies, using Stouffer's Z, to compute the overall significance of the difference between the experimental and control conditions.

Results: Learning Gains

At School A, there was not a significant difference in learning between conditions, i.e. the interaction between condition and time (pre/post), $F(1, 37) = 0.32$, $p = .576$, partial $\eta^2 = .009$. That said, the trend was in the direction of greater learning gains in the ALS (experimental) condition than the New Oriental (control) condition, as shown in Table 1. For the control condition, the pre-test mean = 77.08 and post-test mean = 79.83. The mean score improvement = 2.75 with standard error of 1.49. For the experimental group using ALS's online learning system, the pre-test mean = 78.08 and post-test mean = 82.24. The mean score improvement = 4.16 with standard error of 2.03.

At School B, there was a statistically significant difference in learning between conditions, with students using ALS learning more. For School B, $F(1, 89) = 13.68$, $p < 0.001$, partial $\eta^2 = .133$. For the control group using New Oriental, the pre-test mean = 54.38 and post-test mean = 52.85. The mean score improvement = -1.53 with standard error of 1.35. For the experimental group using ALS, the pre-test mean = 49.36 and post-test mean = 56.09. The mean of score improvement = 6.73 with standard error of 1.80.

At School C, there was not a significant difference in learning between conditions, but there was a non-significant trend towards students using ALS learning more. For School C, $F(1, 88) = 1.90$, $p = .172$, partial $\eta^2 = .021$. For the control group using New Oriental, the pre-test

mean = 64.77 and post-test mean = 65.77. The mean score improvement = 1.00 with standard error of 1.48. For the experimental group using ALS the pre-test mean = 66.48 and post-test mean = 70.57. The mean score improvement = 4.09 with standard error of 1.67.

Table 1

Pre and post-test scores for the two conditions, across the three schools. Standard deviations given in parentheses.

School	Condition	Mean			df	p	F	η^2
		Pre-test	Post-test	Gain				
A	New	77.076 (19.31)	79.83 (17.53)	2.75	1	0.576	0.32	0.009
	ALS	78.08 (18.47)	82.24 (14.32)	4.16				
B	New	54.38 (18.54)	52.85 (20.51)	-1.53	1	<.001	13.68	0.133
	ALS	49.36 (18.01)	56.09 (16.10)	6.73				
C	New	64.77 (14.86)	65.77 (10.10)	1.00	1	0.172	1.890	0.021
	ALS	66.48 (14.26)	70.57 (12.60)	4.09				
				2.75	Stouffer's Z			2.60
					final p			0.01

Considered together using Stouffer's Z, we find that $Z = 2.60$, $p < 0.01$, which indicates that across studies there is a statistically significant effect across the three schools. In specific, students using ALS learned statistically significantly more than students using New Oriental. This pattern was seen in all three schools, but was only statistically significant in one of the three schools.

Analytical Methods: Learning Rate

In this section, we investigate whether students' rate of learning is different between conditions. Some students, for instance, may obtain equal learning using either system, but by avoiding content they already know (due to adaptive choice of concepts for students to cover), they may complete the unit faster. Over the course of the year, this could lead to greater learning by allowing them to focus their time on the concepts they do not yet know, rather than reviewing well-known material.

We operationalize the student's learning rate as the ratio of student's pre-post learning gain to the amount of time they spent using the system (in hours, not rounded), giving a measure of how much improvement the student gained per hour. For instance, a learning rate of 0.04 represents the student gaining 4 percentage points on the test per hour of study. An independent samples t-test is used to test whether there is significant difference of learning rate between two conditions in terms of the two conditions.

As above, we conduct the statistical analysis of the differences in learning rates in these three studies in two steps. First, we conduct a statistical analysis of whether there was a difference in learning rates in each study. Second, we aggregate across the three studies using Stouffer's Z (Mosteller & Bush, 1954) to examine the overall difference in learning rates between conditions.

Results: Learning Rate

At School A, there was not a statistically significant difference in learning rate between conditions, $t(37) = 1.35$, $p = .186$. However, the trend was in the direction of students having a higher learning rate within ALS ($M = .046$, $SD = .095$) than New Oriental ($M = .016$, $SD = .036$).

Within ALS, student study time ranged from 18 minutes to 180 minutes with an average study time of 117.3 minutes. Within New Oriental, student study time ranged from 130 minutes to 180 minutes with an average study time of 172.5 minutes.

At School B, there was a significant difference in learning rate between conditions, $t(89) = 3.59$, $p < 0.001$. Students had a higher learning rate within ALS ($M = .059$, $SD = .107$) than New Oriental ($M = -.016$, $SD = .091$). Within ALS, student study time ranged from 83 minutes to 133 minutes with an average study time of 104.5 minutes. Within New Oriental, student study time ranged from 83 minutes to 133 minutes with an average study time of 104.9 minutes.

At School C, there was not a statistically significant difference in learning rate between conditions, $t(88) = 1.38$, $p = .172$. However, as with School A, the trend was in the direction of students having a higher learning rate within ALS ($M = .040$, $SD = .116$) than New Oriental ($M = .007$, $SD = .107$). Within ALS, student study time ranged from 50 minutes to 129 minutes with an average study time of 97.5 minutes. Within New Oriental, student study time ranged from 50 minutes to 143 minutes with an average study time of 100.2 minutes.

All three schools had trends in the same direction, but these trends were not quite significant in two of the schools. However, when these three studies are integrated together using Stouffer's Z, we find that $Z = 2.95$, $p < 0.01$. The results indicate that, in aggregate, students who use ALS learn more quickly than students using New Oriental. The full pattern of results is given in Table 2.

Table 2

Learning rate in each study condition

School	Condition	Mean	SD	t	df	p	z

A				-1.35	37	.186	-.89
	New Oriental	.016	.036				
	ALS	.046	.096				
B				-3.59	89	<.001	-3.27
	New Oriental	.016	.091				
	ALS	.059	.107				
C				-1.38	88	-0.172	-.95
	New Oriental	.007	.107				
	ALS	.040	.116				
						Stouffer's Z	-2.95
						p	.002

Analytical Methods: Questionnaire Measures

As mentioned above, all students completed a 17-item questionnaire designed by researchers in China after the post-test. Of the 17 items, 5 items are relevant to the topics in this paper:

1. This course is helpful to improve scores on passive voice
2. I think I have a better mastery of passive voice after taking this course
3. It took me less time to learn compared to class at school
4. I think the difficulty of learning content is at the right level
5. I would like to continue using this product to learn other topics on the High School Entrance exam

The first two items assess the student's perception of the system's effectiveness at improving learning. The third item assesses whether the student believes the system is effective at increasing the learning rate. The fourth item assesses whether the student believes the system is effective at tailoring the learning content's difficulty. Finally, the fifth item assesses whether the student would like to continue using the learning content. Each item was given using a Likert scale, from 1 (lowest) to 5 (highest).

In this section, we will investigate, for each of these five items, whether the average response was higher for ALS or for New Oriental. As above, we conduct the statistical analysis of the differences in learning rates in these three studies in two steps. First, we conduct a statistical analysis -- a two-sample two-tailed t-test -- of whether there was a difference in learning rates in each study. Second, we aggregate across the three studies using Stouffer's Z (Mosteller & Bush, 1954) to examine the overall difference in learning rates between conditions.

Results: Questionnaire Measures

Full results are given in Table 3. For item 1, "This course is helpful to improve scores on passive voice", students gave marginally significantly higher responses at School A for ALS (M = 4.29, SD = 0.47) than for New Oriental (M = 3.67, SD = 1.05), $t(27) = 2.03$, $p = .052$. Responses approached marginal significance in the same direction at School B, $t(82) = 1.64$, $p = .105$. The two groups' response averages were identical at School C, $t(86) = 0.00$, $p = 1.00$. When aggregated using Stouffer's Z, ALS had significantly higher responses than New Oriental, $Z = 2.05$, $p = .040$.

For item 2, "I think I have a better mastery of passive voice after taking this course", there was not a statistically significant difference for any of the three schools; for School A, $t(27) = 0.938$, $p = .356$, for School B, $t(82) = 1.57$, $p = .121$, for School C, $t(86) = 0.519$, $p = .605$. However, in all three cases, the trend was towards ALS receiving higher responses than New Oriental. When aggregated using Stouffer's Z, ALS had marginally significantly higher responses than New Oriental, $Z = 1.73$, $p = .084$.

For item 3, "It took me less time to learn compared to class at school", there was a statistically significant difference in favor of ALS ($M = 3.71$, $SD = 1.21$) over New Oriental ($M = 2.93$, $SD = 1.49$) at School B, $t(82) = 2.62$, $p = .010$. A result approaching marginal significance was seen in School A, with ALS ($M = 4.21$, $SD = 0.80$) appearing to perform better than New Oriental ($M = 3.73$, $SD = 0.80$), $t(27) = 1.62$, $p = .117$. In School C, there was a trend towards New Oriental receiving higher scores than ALS, but it was not significant, $t(86) = -0.93$, $p = .354$. When aggregated using Stouffer's Z, ALS had marginally significantly higher responses than New Oriental, $Z = 1.85$, $p = .065$.

For item 4, "I think the difficulty of learning content is at the right level", there was a statistically significant difference in favor of ALS ($M = 4.64$, $SD = 0.50$) over New Oriental ($M = 3.93$, $SD = 0.80$) at School A, $t(27) = 2.85$, $p = .008$. A result approaching marginal significance was seen in School B, $t(82) = 1.55$, $p = .124$. In School C, there was a trend towards ALS receiving higher scores than New Oriental, but it was not significant, $t(86) = 0.69$, $p = .491$. When aggregated using Stouffer's Z, ALS had statistically significantly higher responses than New Oriental, $Z=2.81$, $p=.005$.

Finally, for item 5, "I would like to continue using this product to learn other topics on the High School Entrance exam", there was a marginally significant difference in favor of ALS

(M = 4.50, SD = 0.76) over New Oriental (M = 3.73, SD = 1.28) at School A, $t(27) = 1.94$, p = .062. A result approaching marginal significance was seen in School B, $t(82) = 1.58$, $p = .119$. In School C, there was a trend towards ALS receiving higher scores than New Oriental, but it was not significant, $t(86) = 0.90$, $p = .371$. When aggregated using Stouffer's Z, ALS had statistically significantly higher responses than New Oriental, $Z = 2.49$, $p = .013$.

Table 3

Questionnaire Measures

Item	School	Condition	M	SD	T	p
Q1	A	ALS	4.29	0.47	2.03	0.052
		New Oriental	3.67	1.05		
	B	ALS	4.10	1.02	1.64	0.105
		New Oriental	3.70	1.21		
	C	ALS	4.02	0.98	0.00	1.000
		New Oriental	4.02	0.88		
Q2	A	ALS	4.14	0.86	0.94	0.356
		New Oriental	3.80	1.08		
	B	ALS	3.93	0.93	1.57	0.121
		New Oriental	3.54	1.32		
	C	ALS	4.05	1.06	0.52	0.605
		New Oriental	3.93	1.00		
Q3	A	ALS	4.21	0.80	1.62	0.117
		New Oriental	3.73	0.80		
	B	ALS	3.71	1.21	2.62	0.010
		New Oriental	2.93	1.49		
	C	ALS	3.45	1.19	-0.93	0.354
		New Oriental	3.68	1.10		
Q4	A	ALS	4.64	0.50	2.85	0.008
		New Oriental	3.93	0.80		
	B	ALS	3.85	1.01	1.55	0.124
		New Oriental	3.47	1.26		
	C	ALS	4.02	0.90	0.69	0.491
		New Oriental	3.89	0.95		
Q5	A	ALS	4.50	0.76	1.94	0.062
		New Oriental	3.73	1.28		

B	ALS	3.71	1.42	1.58	0.119
	New Oriental	3.21	1.47		
C	ALS	3.57	1.23	0.90	0.371
	New Oriental	3.34	1.14		

Discussion and Conclusion

In this paper, we have studied the difference between two platforms for online language learning used in China: a non-adaptive platform, New Oriental Online, and an adaptive platform, the Adaptive Learning System (ALS) developed by Classba.com. The two platforms are studied in a set of three experimental comparisons conducted at schools in China. To the best of our knowledge, this study represents the first published investigation into whether adaptive online language learning is more effective than a control condition, within China.

In aggregate, we find that students using ALS learned statistically significantly more than students using New Oriental. This pattern was seen in all three studies, but was only statistically significant in one of the three studies when taken individually. We also find that, in aggregate, students who use ALS learn more quickly than students using New Oriental. Again, this pattern was seen in all three studies, but was only statistically significant in one of the three studies, when taken individually. This finding suggests that, on the whole, adaptive learning features have the potential to improve student outcomes – helping them learn better and faster than non-adaptive approaches.

In terms of student attitudes, we find that students reported better attitudes towards ALS than New Oriental for five different measures, although again these results were not statistically significant in every individual study. This finding matches previous findings from single-condition studies that online language learning is associated with positive attitudes among students within China (孙秋丹, 张健, & 董哲. 2003; Zhao, 2015).

Overall, these results represent evidence that the innovation presented in ALS -- adaptive learning tailored to learning paths, to select the optimal content for each student -- is generally more effective and preferred by students in China.

Of course, the results of just three studies cannot be considered conclusive. These studies, while distinct in some ways in terms of their design and implementation, compared between the same two platforms. While these platforms were broadly similar in their functionality (excluding adaptivity), the two conditions did not use the same videos, for instance, creating some confounding in which design features led to the better results for ALS. It is possible that even had ALS not been adaptive, the difference between videos might have resulted in better learning outcomes. At this same time, this study's findings accord with mounting evidence that adaptive learning systems are more effective than other instructional approaches, in other countries, for a range of topics (Abrami et al., 2015; Mitrovic, Martin, & Mayo, 2002; Pane et al., 2014; Tseng, Chu, Hwang, & Tsai, 2008).

In addition, not all regions of China were included in the studies reported within this paper, as doing so would be a major endeavor. The studies also did not explicitly include all the demographic groupings in China. Finally, only one educational topic was included in this study. As such, these results must be considered as indicative but still preliminary.

Nonetheless, we view this study as an important step towards implementing the same type of evidence base for curricular decision-making in China as is currently seen in national projects like the What Works Clearinghouse. Through the implementation of more randomized controlled experiments in education, we can better understand what designs and types of adaptation are effective for learners in China, and what curricula are most effective at enhancing learning outcomes.

Acknowledgements

The authors would like to thank the principals and teachers of the three participating middle schools for their support to this research study. The authors would also like to thank Dataway for their professional consulting on the design and implementation of the study, and Junyi Li and Yang Jiang for assistance in literature review. Finally, the authors would like to thank Alexis Andres for assistance in paper formatting and preparation.

References

Abrami, P., Borohkovski, E., & Lysenko, L. (2015). The effects of ABRACADABRA on reading outcomes: A meta-analysis of applied field research. *Journal of Interactive Learning Research, 26*(4), 337-367.

Abreu-Ellis, C., Ellis, J. B., Carle, A., Blevens, J., Decker, A., Carvalho, L., & Macedo, P. (2013). Language learning: The merge of Teletandem and Web 2.0 tools. *Journal of Interactive Learning Research*, *24*(4), 353-369.

Beatty, K. (2013). *Teaching & Researching: Computer-Assisted Language Learning*. Routledge.

Chuang, F. Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora*, *1*(2), 251-271.

Doignon, J. P., & Falmagne, J. C. (2012). *Knowledge spaces*. Springer Science & Business Media.

Kasteen, J. (2014). Global trends in foreign language demand and proficiency. In *ICEF Monitor*. Retrieved from http://studenttravelplanningguide.com/global-trends-in-foreign-language-demand-and-proficiency/

Kulik, J.A., & Fletcher, J.D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytics review. *Review of Educational Research, 86*(1), 42-78.

Kroeze, K., Hyatt, K., & Lambert, C. (2015) Brain Gym: Let the user beware. *Journal of Interactive Learning Research, 26*(4), 395-401.

Levis, J. M., Sonsaat, S., Link, S., & Barriuso, T. A. (2016). Native and nonnative teachers of L2 pronunciation: Effects on learner performance. *TESOL Quarterly*. *1-38 (Early View).*

Li, E. (2016). The Study on Higher Vocational English O2O Teaching Experiment from the Perspective of MOOC. *Journal of HUBEI Correspondence University, 29*, 136-137.

Lipsey, M.W., Wilson, D.B. (2001). *Practical Meta-Analysis.* Thousand Oaks, CA: Sage Publications.

Macaro, E., Handley, Z., & Walter, C. (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching*, *45*(01), 1-43.

Marty, F (1981). Reflections on the use of computers in second language acquisition. *System*, *9*(2), 85–98.

Mitrovic, A., Martin, B., & Mayo, M. (2002). Using evaluation to shape ITS design: Results and experiences with SQL-Tutor. *User Modeling and User-Adapted Interaction*, *12*(2), 243-279.

Mosteller, R., L., Bush, R.R. (1954). *Selected quantitative techniques*. In G. Lindzey (Ed.), *Handbook of Social Psychology,* vol. 1. Cambridge, MA: Addison-Wesley.

Ogan, A., Walker, E., Baker, R.S.J.d., de Carvalho, A., Laurentino, T., Rebolledo-Mendez, G., Castro, M.J. (2012). Collaboration in cognitive tutor use in Latin America: Field study and design recommendations. *Proceedings of ACM SIGCHI: Computer-Human Interaction*, 1381-1390.

Orwin, R.G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157-159.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127-144.

Rodrigo, M.M.T., Baker, R.S.J.d., Rossi, L. (2013). Student off-task behavior in computer-based learning in the Philippines: Comparison to prior research in the USA. *Teachers College Record*, *115* (10), 1-27.

Roschelle, J., Feng, M., Murphy, R.F., Mason, C.A. (2016) Online Mathematics Homework Increases Student Achievement. *AERA Open, 2* (4).

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. Boston, MA: McGraw-Hill Humanities Social.

Sun, Q., Zhang, J., Dong, Z. (2003) Research into the teaching of professional English: use of computers and internet. *Foreign Language World* (外语界) *(2),* 28-33.

Tseng, J. C., Chu, H. C., Hwang, G. J., & Tsai, C. C. (2008). Development of an adaptive learning system with two sources of personalization information. *Computers & Education*, *51*(2), 776-786.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.

von Ahn, L. (2013). Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent User Interfaces,* 1-2. ACM.

Wang, S., Liu, J., & Fu, W. (2015). A Study on the Content-based college ESP Instruction under the Background of MOOC Era. *China Educational Technology, 2015*, 98-120.

Zhao, N-s. (2015) A Study on the Effects of Online Interaction for the EFL Learners' Writing Motivation. *Overseas English, 2*, 133-136.