TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

August 2016 – Volume 20, Number 2

Measuring the Impact of Language-Learning Software on Test Performance of Chinese Learners of English

Justin Nicholes
Indiana University of Pennsylvania, USA
<j.nicholes@iup.edu>

Abstract

This classroom quasi-experiment aimed to learn if and to what degree supplementing classroom instruction with Rosetta Stone (RS), Tell Me More (TMM), Memrise (MEM), or ESL WOW (WOW) impacted high-stakes English test performance in areas of university-level writing, reading, speaking, listening, and grammar. Seventy-eight (N = 78) Chinese learners of English in a cross-border higher education program with a Midwestern U.S. state comprehensive university made up the participants. From the pretest/posttest repeated-measures design, quantitative results included (1) no practice-and-drill software group (RS, TMM, and MEM) significantly outperforming the reference software group (WOW) although statistically significant improvements appeared among subsections of the test and among test totals; (2) no significant correlation between number of hours spent using technology and posttest scores; and (3) statistically significant differences between female and male participants in all groups in the areas of pretest scores, posttest scores, and total number of hours using language-learning software over the study's 15-week period. These findings offer insight into supplementing language classroom instruction with language learning software in Mainland China, with implications for language teachers across Asia.

235,597 Chinese students (28.7 percent of all international pupils) enrolled in U.S. universities for the 2012-2013 academic year, a 21 percent increase from the previous academic year (Institute of International Education, 2013). Increasingly, Chinese learners begin or continue U.S.-based higher education while never having to leave China. This, along with available literature on the interrelations of cross-border higher education (CBHE), English performance on high-stakes standardized tests, and students' money-making potential in China, explains the need for sustained attention Chinese learners deserve from institutions of higher learning.

In China, a competitive job market underscores the importance of higher education and particularly of English test performance. Employers in China view a job applicant's academic-performance record as a device for signaling ability and screening out

unqualified applicants (Wang & Morgan, 2009), and young workers in China with advanced skills have experienced jobless rates of approximately 4.1% in a context where roughly 8% are reported as unemployed (Orlik, 2012). After the economic crisis of 2009, jobs in China became harder to find for university graduates (Qi, 2011), and the reputation of the university related to the amount of money graduates with bachelor degrees tended to earn, correlating with 25-30% higher wages for students with degrees from top-100 Chinese universities and with 15-20% higher wages for graduates from top-200 schools (Hartog, Sun, & Ding, 2010). Students from the most prestigious Chinese schools settled least frequently for jobs they did not want and suffered least frequently from dissatisfaction from over-education (Li, Zhao, & Tian, 2010). Chinese learners who cannot matriculate into top-tier universities, however, can still benefit from performance on high-stakes standardized English-language tests. Good grades in conjunction with "a National Standard English Certificate" contribute to Chinese graduates finding the jobs they want and making more money (Zhao, 2009).

For years, teachers, researchers, administrators, and government officials in China have looked to computer-assisted language learning (CALL) to boost learners' performance of English on standardized tests. They have pointed out the growing interest of CALL and the need for teachers to develop learners' English in response to growing access to technology in China (Li, 2007), with Ruan (2008) specifically calling for classroom-based research to explore local issues related to Chinese learners of English and CALL.

This classroom-based quasi-experiment used a quantitative repeated-measures design to address a local problem at the research site, hoping to skirt the unwitting duplication of effort that Shield (2009) warned against when hypotheses and designs previously tested on obsolete technology reappear in publications on newer, more popular technology. Specifically, this study aimed to learn if and to what degree supplementing classroom instruction with Rosetta Stone (RS), Tell Me More (TMM), Memrise (MEM), or ESL WOW (WOW) impacted English performance on a high-stakes standardized test meant to measure writing, reading, speaking, listening, and grammar of Chinese learners of English in a CBHE context in Mainland China.

So far, researchers have focused only on RS and TMM. Peer-reviewed studies criticizing RS from theoretical standpoints have identified erroneous theoretical underpinnings, culturally inappropriate images, and little connection to learners' lives (DeWaard, 2013; Lopez-Lopez, 2012). Among the few available participant-based studies, Nielson's (2011) study explored how adult learners used both RS and TMM and, to the chagrin of the researcher, found huge attrition among participants and negligible benefit for learners. Meanwhile, Demir and Korkmaz (2013) reported significant changes among pre- and posttest performance measures for listening and speaking after 21 hours of RS use versus traditional teaching-English-as-a-foreign-language (TEFL) classroom instruction, while a participant-based study on learner perceptions of TMM reported that learners perceived the software as helpful for learning English (Hashim & Yunus, 2010).

Overall, deficiencies in the evidence appear on two levels. From a research-based point of view, few participant-based studies have looked at RS or TMM, while none have looked at MEM or WOW. Of the existing research on RS and TMM, seemingly nobody has situated research objectives into an experimental design or had access to this present study's kind

of research site (Chinese CBHE) or participants. In addition, although Hashim and Yunus (2010) measured learner perceptions of TMM, the study failed to report how much participants used the technology during the study. Deficiencies in the evidence appear also from a practice-based point of view. While previous research has looked at RS and TMM as stand-alone packages, nobody has studied the impact of these programs as supplements to classroom instruction or in comparison with the impact of alternative software.

In light of these areas where knowledge can be built upon, the following research question guided the present study: How do Rosetta Stone, Tell Me More, Memrise, and ESL WOW impact the English performance on a high-stakes standardized test meant to measure reading, writing, listening, speaking, and grammar?

Literature Review

To justify the need for this study, the following literature review reiterates and further details places of potential growth in knowledge related to Chinese learners of English and to the language-learning programs used here.

Rosetta Stone, Tell Me More, Memrise, and ESL WOW

Few peer-reviewed scholarly articles have explored RS and TMM, while no article has examined either MEM or WOW. Using Cummins, Brown, and Sayers's (2007) framework for evaluating software in schools, Lopez-Lopez (2012) pointed out perceived shortcomings of RS, including, (a) a lack of opportunities for critical thinking or cognitive challenge; (b) a failure to clarify deep connections between vocabulary, grammar, and the learners' lives; (c) a limit on collaborative inquiry in the online games and practice activities; (d) a gap in any kind of engaging reading or writing practice across the curriculum; and (e) no chance to develop effective reading or writing strategies. Still, the program appeared to offer opportunities for "effective involvement and identity investment" through collaboration with other students and instructors online (Lopez-Lopez, 2012, p. 7). DeWaard (2013) likewise evaluated RS in light of the company's push to enter academia and concluded that the software package was "not a viable option for language learning" (p. 61), citing unclear theoretical underpinnings, culturally inappropriate images, and nonstandard proficiency levels. Nielson (2011) explored how adult learners used both RS and TMM and, consistent with literature on the ineffectiveness of self-study alone (Benson, 2007; Holec, 1981), found significant attrition: Of the more than 300 participants, only 5 completed the full study protocol, with half of the participants never accessing the software once and only a handful benefiting from the study. Hashim and Yunus (2010) surveyed the perceptions learners of English held regarding the usefulness of TMM; most participants thought the tool was easy to use, almost all felt it helped them improve their English performance, and most thought it was suitable for language learning. Still, the survey design failed to either measure or report how much participants used the software or whether participants showed improvement on performance measures. In contrast, this present study sought to measure both the amount of time learners used the software and the extent to which participants' mean scores on pre- and posttests changed over the duration of the study.

Chinese English Language Learners and CALL

A significant amount of previous research on Chinese English language learners (ELLs) has explained learner tendencies and representative English-language settings in China. Previous studies have emphasized the importance of instrumental motivation for Chinese ELLs, with learners' goals including scoring favorably on high-stakes exams, winning scholarships, receiving praise from teachers, earning higher grades, or landing good jobs (Li, 2014; Liu, 2012; Ning & Hornby, 2013; Peng & Woodrow, 2010; Zhang & Guo, 2012; Zheng, 2012). Earlier research also has highlighted anxiety as affecting Chinese ELLs motivated learning behavior (Li, 2014) and willingness to speak English in class (Li, 2014; Liu, 2006; Liu & Jackson, 2011). Though some research indicated that female Chinese participants tended to report higher motivation to learn English than males (Lamb, 2004; Liu, 2009, 2012; Yang, Liu, & Wu, 2010), biological sex of Chinese ELLs did not seem to predict significant differences in English proficiency or language-learning strategy use in other research (Nisbet, Tindall, & Arroyo, 2005). As Zhou and Xu (2012) found, although currently in China adolescents have more freedom to choose college majors than earlier generations, those whose universities selected majors for them tended to have more negative attitudes toward learning, less inclination to pick up new learning strategies, and poorer academic performance. This lack of choice may extend to Chinese teachers as well. You (2004) found overworked, institutionally-restricted Chinese English writing instructors typically followed a current-traditional approach, emphasizing correct form over thought development to not prepare students for communication, but to ready learners for high-stakes exams.

Along these lines, Chinese ELLs have been described as remaining accustomed to test- and teacher-centered classrooms where lectures dominate (Li, 2014; Lu, Li, & Du, 2009) and where a focus on receptive knowledge of English vocabulary and prescriptive grammar sometimes prevents active, productive use or communication (Ma, 2012; Peng & Woodrow, 2010; Zheng, 2012). As a result of compulsory learning of English, Chinese learners may not enjoy learning English (Ning & Hornby, 2013). In addition, other research has pointed out that, in spite of years of studying English, many Chinese students perform poorly on speaking measures (Peng, 2014). With these challenges in mind, many have looked to CALL as a way to motivate students and boost Chinese ELLs' performance on tests and elsewhere.

Lu, Li, and Du (2009) discussed the Chinese Department of Higher Education's call for more student-centered, communicative language teaching (CLT) approaches in English language classes, especially through the incorporation of CALL. A perceived lack of speaking and listening skills inspired these directives, with Liao (2004) arguing that the government's position of embracing CLT would "bring about a positive effect on English teaching and learning" (p. 272). Still, Niu-Cooper (2012) noted challenges regarding change in China, including Chinese teachers' habitually teacher-centered styles and crowded classrooms that ended up restricting pair and group work. Regarding CLT-guided CALL, Li (2007) listed more challenges to change in Chinese English-language-learning contexts, such as scarcity of reliable hardware and software, teachers' lack of training in CALL who remained reluctant to change teaching styles, and teachers' uncertainty about how to use CALL materials.

Previous CALL research in China has implored teachers of English to develop relevant skills in teaching with technology (Li, 2007). While warning researchers in China not to carry out CALL-related studies simply to tout a popular theory, Ruan (2008) called for classroom-based designs to explore how best to use CALL in existing classrooms. Past studies have illustrated significant gains in speaking performance, including varied vocabulary and more complex syntax, after e-learning treatment in English-language speaking classes in China (Shen & Suwanthep, 2011). Zou (2011) also found that, given teacher training and reliable hardware, CALL in China fostered learner autonomy.

Justification for the Present Study

The above literature justifies the present study in two main ways: (a) This study sought to give much-needed attention to popular language-learning programs that are increasingly entering academia and, also, that have attracted little to no attention in the way of participant-based studies, and (b) this study explored whether supplementing classroom instruction with popular language-learning software significantly impacted language performance on a high-stakes language test, offering practical guidance for administrators and instructors involved with and invested in the research site.

Methods

Research Question

The researcher investigated the following research question: How does Rosetta Stone, Tell Me More, Memrise, and ESL WOW impact the English performance on a high-stakes standardized test meant to measure reading, writing, listening, speaking, and grammar?

Participants

Seventy-eight (N = 78; 43 females, 35 males) Chinese ELLs between the ages of 20-22 studying at a private Chinese college in a cross-border program with a Midwestern U.S. state-comprehensive university (SCU) made up the participants in this study. The researcher used a nonprobability sampling approach of convenience representative of classroom-based research before the semester began and randomly identified language-learning software for each of his four sections of English writing: RS: n = 19 (5 females, 14 males); TMM: n = 19 (14 females, 5 males); MEM: n = 21 (12 females, 9 males); and WOW (n = 19 (13 females, 6 males). All participants spoke Mandarin Chinese. Students were placed into the cross-border program based on scores earned on the Chinese university's entrance exam. The researcher's university Internal Review Board approved the research and data-collection measures. Permission was obtained from the Chinese partner university, and all participants gave informed consent in Mandarin.

Materials

Rosetta Stone. Stoltzfus (1997) identified the theoretical underpinnings of RS as the comprehension or natural theory, in which second language acquisition (SLA) starts with listening comprehension, mimicking the process of first language acquisition (FLA). Lopez-Lopez (2012) noted that the program melded text, images, and sound at student-appropriate levels while describing "immersion" in RS as involving "intuition, interaction, and instruction" (p. 2). Accordingly, learners view an image with English text and then hear and are prompted to repeat or explain the stimulus into a microphone, allowing a

speech-recognition tool to measure comprehensibility. Drawing on personal experience in using the package to learn German, Hiebert (2012) wrote that RS supported SLA more effectively than the text-image coupling in children's books by providing "critical and consistent data without substantial amounts of diverting information" (p. 291). Learners can chat with other RS learners who happen to be online and can also eventually talk with an RS tutor.

Tell Me More. TMM includes more drilling in grammar and vocabulary than RS and also seems to feature a more sensitive speech-recognition tool. TMM involves fill-in-the-blank exercises and puzzles and gives learners the choice of focusing on everyday English or on business/professional English. In a description of TMM, Brynko (2008) noted, "six workshop activities: culture (geography lessons and city tours), written (practice your prose), vocabulary and grammar (fine-tune your skills), oral (speak the language), lessons (activities from matching words to multiple choice), and dialogue (using speech recognition)" (p. 41). During the semester in which the researcher was carrying out the project, RS acquired TMM.

Memrise. MEM is a free online learning tool with user-created classes (memrise.com). In language classes, learners mostly match words, phrases, or sentences with definitions or translations. If the learner forgets the meaning of the target form, he or she may review an image that the user had earlier selected or created to correspond with that target form. Learners also may hear voice recordings of English speakers repeating target forms. Learners earn points for completing levels, and usernames get posted on weekly, monthly, and all-time leaderboards in real time in competition with users worldwide. Learners produce language by typing answers into blanks while a timer clicks toward zero.

ESL Writing Online Workshop (ESL WOW). Essentially a reference website explaining the process of writing, WOW features animations of a student getting help from a writing-center tutor (esl-wow.org). Unlike the other practice-and-drill tools above, this online tool did not offer self-paced routes of learning or any recorded assessment. Instead, learners watched animations and listened to dialogs centered on English composition topics, such as "Getting Ready to Write," "Developing Your Ideas," "Revising Your Work," and "Editing and Polishing." Learners had no opportunities to speak or write language or to engage with or manipulate the videos aside from pausing or replaying.

Assessments. A version of a placement exam used at the researcher's U.S. university served as the pre- and posttest. The exam included, (a) two 30-minute reading sections with a total of 15 questions related to two 300-word passages; (b) a 45-minute writing section that asked participants to compose an argument in the form of an essay; (c) a 30-minute listening section of 20 questions in response to a 3-minute audio dialog; (d) an approximately 5-to-7-minute speaking section of a one-on-one conversation with the researcher; and (e) a 30-minute grammar section of 50 discrete-item and integrative questions.

To measure amount of time learners spent using language-learning software each week, the researcher made use of built-in assessments in RS and TMM, as well as the university's learning management system (LMS), through which participants gained access to the software.

Procedures

In the previous semester, the researcher piloted the quasi-experiment. At the start of the 15-week semester that made up this study's duration, participants met the researcher in a computer lab. Those who gave informed consent (N = 78) completed surveys as well as pretests that measured reading, writing, speaking, listening, and grammar. Throughout the semester, learners were asked to use the software at least three hours per week and to record progress in weekly, LMS-hosted electronic journals. At the end of the semester, participants retook surveys and completed the posttest. The participants were debriefed, thanked, and added to an email list that would let them know when the study might be either published or otherwise reported.

Results

To answer the research question, the researcher ran, (a) outlier tests, Shapiro-Wilk tests of normal distribution, and Levene's test for homogeneity of variances; (b) Welch ANOVA and Games-Howell post hoc tests to check for significant variance between groups; (c) paired t-tests on pre- and posttests, followed up by nonparametric Wilcoxon's tests; and finally, (d) Pearson product-moment correlation tests on the number of hours participants used language-learning software and posttest scores. The researcher used IBM SPSS Statistics version 21 to analyze data.

Outliers, Normal Distribution, and Homogeneity of Variances

Descriptive results showed outliers in the RS group. Cook's distance regression analysis (D) indicated that RS observation 1 held (D = 0.101) and RS observation 18 showed (D = 0.042). The researcher did not eliminate outliers' scores from the data set, opting instead for robust statistical analyses and noting this limitation common in classroom-based research and convenience sampling. Shapiro-Wilk's test for normal distribution showed that, overall, groups' scores demonstrated normal distribution at the $p > 0.05$ level: RS (p = 0.75), TMM (p = 0.782), MEM (p = 0.268), and WOW (p = 0.092); however, evidence of normal distribution appeared to be lacking in sections of the tests for some groups. Finally, Levene's test for homogeneity of variances found a lack of homogeneity at the $p >$ 0.05 level between the variances in the population (p = 0.028), indicating that variability of scores did not remain equal across the four groups and that equal variances among groups' pretest scores could not be assumed.

In conclusion, (a) significant outliers appeared in the RS group's pretest but remained in later analyses (accounted for with robust tests); (b) significant normal distribution appeared among groups on pretests though a lack of normal distribution appeared among sections of the test for some groups; and (c) variability of scores did not remain equal across the groups. In response to these initial findings on pretest scores, the researcher selected the Welch ANOVA instead of one-way ANOVA test, Games-Howell post hoc instead of Tukey's post hoc, and followed up t-tests with Wilcoxon's nonparametric test.

Welch ANOVA and Games-Howell Test on Pretest Scores

Regarding the extent groups' performance on pretests resembled each other, Welch ANOVA tests showed that, although groups' pretest scores exhibited no significant difference among one another in reading, listening, and grammar, significant differences

appeared at the $p < 0.05$ level in writing ($p = 0.001$), speaking ($p < 0.001$), and totals ($p < 0.001$). Games-Howell post hoc tests illuminated which groups differed and to what degree. RS pre-writing (M = 54.21, SD = 11.04) proved significantly lower than MEM pre-writing (M = 64.86, SD = 6.89) with significance at the $p < 0.05$ level of ($p = 0.006$) and also proved significantly lower than WOW pre-writing (M = 69.16, SD = 8.80) ($p < 0.001$). RS pre-speaking (M = 50.00, SD = 12.25) proved significantly lower than TMM pre-speaking (M = 82.90, SD = 12.62) ($p < 0.001$), WOW pre-speaking (M = 73.16, SD = 13.66) ($p < 0.001$), and MEM pre-speaking (M = 64.29, SD = 8.26) ($p = 0.001$). At the same time, MEM pre-speaking (M = 64.29, SD = 8.26) proved significantly lower than TMM pre-speaking (M = 82.90, SD = 12.62) ($p < 0.001$). Regarding total scores, RS pretest (M = 53.00, SD = 8.56) proved significantly lower than both TMM pretest (M = 65.84, SD = 7.65) ($p < 0.001$) and WOW pretest (M = 62.47, SD = 10.47) ($p = 0.022$) but not significantly lower than MEM pretest (M = 57.71, SD = 5.41). In turn, MEM pretest (M = 57.71, SD = 5.41) proved significantly lower than TMM pretest (M = 65.84, SD = 7.65) ($p = 0.003$).

In summary, the groups showed no significant variance regarding their reading, listening, and grammar pretest scores. Still, significant variance appeared elsewhere: (a) RS scored significantly lower than MEM and WOW on writing, significantly lower than all other groups on speaking, and significantly lower than TMM and WOW in total scores; (b) MEM, in turn, scored significantly lower than TMM on speaking and significantly lower than TMM on the pretest total. Pretest performance before treatment appeared in this order: TMM (65.8%), WOW (62.5%), MEM (57.7%), and RS (53%).

Finally, Welch ANOVA tests indicated significant difference between pretest means of female participants (M = 62, SD = 8.23) and pretest means of male participants (M = 56.74, SD = 10), with significance at the $p < 0.05$ level of ($p = 0.015$).

Paired Samples Analyses

To measure for difference between pre- and posttest means, the researcher ran analyses on paired samples. Table 1 below describes paired samples from pre- and posttests for all groups.

Table 1. ***Groups' Paired Samples Statistics***

Pairs	Mean	N	SD	Std. Error Mean
RS pre-Reading	45.895	19	16.248	3.728
RS post-Reading	44.842	19	18.446	4.232
TMM pre-Reading	49.105	19	13.212	3.031
TMM post-Reading	55.158	19	12.562	2.882
MEM pre-Reading	39.857	21	14.578	3.181
MEM post-Reading	44.381	21	14.596	3.185
WOW pre-Reading	48.316	19	19.096	4.381
WOW post-Reading	47.421	19	21.122	4.846

Pairs	Mean	N	SD	Std. Error Mean
RS pre-Writing	54.210*	19	11.038	2.532
RS post-Writing	63.260*	19	11.416	2.619
TMM pre-Writing	63.632*	19	10.935	2.509
TMM post-Writing	68.211*	19	11.989	2.750
MEM pre-Writing	64.857*	21	6.887	1.503
MEM post-Writing	72.429*	21	6.454	1.408
WOW pre-Writing	69.158*	19	8.802	2.019
WOW post-Writing	72.053*	19	6.329	1.452
RS pre-Listening	54.470	19	12.681	2.909
RS post-Listening	49.740	19	17.910	4.109
TMM pre-Listening	63.421	19	13.443	3.084
TMM post-Listening	65.790	19	11.698	2.684
MEM pre-Listening	53.619	21	11.205	2.445
MEM post-Listening	54.048	21	10.796	2.356
WOW pre-Listening	52.632–	19	17.270	3.962
WOW post-Listening	59.474–	19	15.714	3.605
RS pre-Speaking	50.000	19	12.247	2.810
RS post-Speaking	51.580	19	12.478	2.863
TMM pre-Speaking	82.895	19	12.618	2.895
TMM post-Speaking	81.579	19	14.342	3.290
MEM pre-Speaking	64.286*	21	8.259	1.802
MEM post-Speaking	68.810*	21	9.605	2.096
WOW pre-Speaking	73.158*	19	13.664	3.135
WOW post-Speaking	77.368*	19	11.828	2.714
RS pre-Grammar	60.790~	19	13.811	3.168
RS post-Grammar	44.840~	19	18.446	4.232
TMM pre-Grammar	70.000	19	12.065	2.768
TMM post-Grammar	69.790	19	11.703	2.685
MEM pre-Grammar	65.952	21	12.343	2.694
MEM post-Grammar	68.381	21	10.509	2.293

Pairs	Mean	N	SD	Std. Error Mean
WOW pre-Grammar	69.105	19	11.907	2.732
WOW post-Grammar	71.368	19	11.908	2.732
RS pretest TOTAL	53.000	19	8.564	1.965
RS posttest TOTAL	54.620	19	11.400	2.615
TMM pretest TOTAL	65.842	19	7.654	1.756
TMM posttest TOTAL	68.000	19	8.901	2.042
MEM pretest TOTAL	57.714*	21	5.414	1.182
MEM posttest TOTAL	61.571*	21	5.250	1.146
WOW pretest TOTAL	62.474*	19	10.474	2.403
WOW posttest TOTAL	65.526*	19	9.571	2.196
*Significantly better score (paired *t*-test and Wilcoxon's test) –Significant on paired *t*-test but nominally insignificant on Wilcoxon's test ~Significantly worse score				

Table 1 shows the extent to which all groups improved on posttests. Closer analysis, though, determined whether changes between pre- and posttest scores proved significant (see Table 2).

Table 2. ***Groups' Paired Samples t-tests***

	Paired Differences					t	df	Sig. (2-tailed) (*p<0.05)
Pairs	Mean	SD	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
RS Reading	1.053	16.026	3.677	-6.672	8.777	.286	18	.778
TMM Reading	-6.053	15.342	3.520	-13.447	1.342	-1.720	18	.103
MEM Reading	-4.524	16.603	3.623	-12.081	3.034	-1.249	20	.226
WOW Reading	.895	11.638	2.670	-4.714	6.504	.335	18	.741
RS Writing	-9.053*	8.442	1.937	-13.122	-4.984	-4.674	18	.000*

	Paired Differences					t	df	Sig. (2-tailed) (*p<0.05)
Pairs	Mean	SD	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TMM Writing	-4.579*	5.368	1.231	-7.166	-1.992	-3.718	18	.002*
MEM Writing	-7.571*	7.763	1.694	-11.105	-4.038	-4.470	20	.000*
WOW Writing	-2.895*	4.795	1.100	-5.206	-.584	-2.632	18	.017*
RS Listening	4.737	18.141	4.162	-4.007	13.480	1.138	18	.270
TMM Listening	-2.368	9.771	2.242	-7.078	2.341	-1.057	18	.305
MEM Listening	-.429	12.464	2.720	-6.102	5.245	-.158	20	.876
WOW Listening	-6.842*	14.163	3.250	-13.668	-.016	-2.106	18	.050*
RS Speaking	-1.579	6.248	1.433	-4.590	1.432	-1.102	18	.285
TMM Speaking	1.316	9.696	2.224	-3.357	5.989	.592	18	.562
MEM Speaking	-4.524*	5.456	1.191	-7.007	-2.041	-3.800	20	.001*
WOW Speaking	-4.211*	7.502	1.721	-7.827	-.595	-2.446	18	.025*
RS Grammar	15.947*	20.791	4.770	5.926	25.968	3.343	18	.004*
TMM Grammar	.211	7.656	1.757	-3.480	3.901	.120	18	.906
MEM Grammar	-2.429	9.796	2.138	-6.888	2.030	-1.136	20	.269
WOW Grammar	-2.263	5.646	1.295	-4.984	.458	-1.747	18	.098
RS TOTAL	-1.620	4.910	1.126	-3.986	.746	-1.438	18	.168

	Paired Differences					t	df	Sig. (2-tailed) (*p<0.05)
Pairs	Mean	SD	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TMM TOTAL	-2.158	5.853	1.343	-4.979	.663	-1.607	18	.125
MEM TOTAL	-3.857*	4.305	.939	-5.817	-1.898	-4.106	20	.001*
WOW TOTAL	-3.053*	3.613	.829	-4.794	-1.311	-3.683	18	.002*

Summary of Significant Improvement

Reading. Analysis of pre- and posttest means revealed that no group significantly improved its reading scores; however, statistically insignificant improvements appeared between all groups' pre- and posttest mean scores. Wilcoxon's signed rank test confirmed a lack of statistically significant improvement.

Writing. Analysis of pre- and posttest means revealed that all groups significantly improved at the $p < 0.05$ level in writing. RS pre-writing (M = 54.21, SD = 11.04) differed significantly from RS post-writing (M = 63.26, SD = 11.42); t(18)=-4.67, $p < 0.001$. MEM pre-writing (M = 64.86, SD = 6.89) differed significantly from MEM post-writing (M = 72.43, SD = 6.45); t(20)=-4.47, $p < 0.001$. TMM pre-writing (M = 63.63, SD = 10.94) differed significantly from TMM post-writing (M = 68.21, SD = 11.99); t(18)=-3.72, $p = 0.002$. Finally, WOW pre-writing (M = 69.16, SD = 8.80) differed significantly from WOW post-writing (M = 72.05, SD = 6.33); t(18)=-2.63, $p = 0.017$. Wilcoxon's signed rank test confirmed significant improvements at the $p < 0.05$ level: RS ($p = 0.001$), MEM ($p = 0.001$), TMM ($p = 0.006$), and WOW ($p = 0.022$).

Listening. Analysis of pre- and posttest means revealed that only the WOW group improved its listening score. WOW pre-listening (M = 52.63, SD=17.27) differed significantly from WOW post-listening (M = 59.47, SD = 15.71); t(18)=-2.11, $p = 0.05$; however, Wilcoxon's signed rank test showed nominally insignificant change at the $p < 0.05$ level ($p = 0.051$). TMM and MEM both made statistically insignificant improvements while RS performed worse on the listening posttest, though not significantly so. Wilcoxon's signed rank test confirmed these findings.

Speaking. Analysis of pre- and posttest means revealed that only MEM and WOW significantly improved on speaking. MEM pre-speaking (M = 64.29, SD = 8.26) differed significantly from MEM post-speaking (M = 68.81, SD = 9.61); t(20)=-3.80, $p = 0.001$. WOW pre-speaking (M = 73.16, SD = 13.66) differed significantly from WOW post-speaking (M = 77.37, SD = 11.83); t(18)=-2.45, $p = 0.025$. Statistically insignificant improvements appeared from RS and TMM. Wilcoxon's signed rank test confirmed

significant improvements at the $p < 0.05$ level, MEM ($p = 0.002$) and WOW ($p = 0.035$), as well as statistically insignificant improvements of RS and TMM.

Grammar. Analysis of pre- and posttest means revealed that no group significantly improved its grammar scores; however, statistically insignificant improvements appeared from MEM, WOW, and TMM. RS scored significantly lower on grammar posttests, with RS pre-grammar (M = 60.79, SD = 13.81) differing significantly from RS post-grammar (M = 44.84, SD = 18.45); t(18)=3.34, $p = 0.004$. Reasons for this probably include, (a) RS group's starting out at a lower point at the beginning of the semester; (b) RS group's spending significantly less time using the technology compared with other groups (to be discussed in detail below); and (c) RS group's higher percentage of male participants, who scored significantly lower than female participants in the study (to be discussed in detail below). Wilcoxon's signed rank test confirmed the RS group's significant change on grammar performance at the $p < 0.05$ level ($p = 0.005$).

Total scores. Analysis of pre- and posttest means revealed significant improvement from MEM and WOW. MEM pretest (M = 57.71, SD = 5.41) differed significantly from MEM posttest (M = 61.57, SD = 5.25); t(20)=-4.11, $p = 0.001$. WOW pretest (M = 62.47, SD = 10.47) differed significantly from WOW posttest (M = 65.53, SD = 9.57); t(18)= -3.68, $p = 0.002$. Insignificant improvements came from RS and TMM. Wilcoxon's signed rank test confirmed significant changes at the $p < 0.05$ level, MEM ($p = 0.001$) and WOW ($p = 0.003$), as well as statistically insignificant improvements of RS and TMM.

Welch ANOVA tests indicated significant difference between posttest means of female participants (M = 65.05, SD = 8.57) and posttest means of male participants (M = 58.99, SD = 11.08), with significance at the $p < 0.05$ level of ($p = 0.008$).

Posttest Scores and Time Spent Using Language-Learning Software

To understand how time using language-learning software related to participants' test scores, the researcher checked for normal distribution, checked for homogeneity of variances, ran one-way ANOVA tests, ran Tukey's post hoc test, and finally ran Pearson product-moment correlation tests.

Shapiro-Wilk tests showed groups' means of hours using technology exhibited significant normal distribution at the $p > 0.05$ level: RS ($p = 0.714$), TMM ($p = 0.331$), MEM ($p = 0.703$), WOW ($p = 0.188$). Next, Levene's test of homogeneity of variances found significant homogeneity at the $p > 0.05$ level ($p = 0.112$). This measurement showed that variability of mean hours of using language-learning software remained equal across the four groups and that significantly equal variances could be assumed. Table 3 describes hours participants spent using language-learning software.

Table 3. ***Hours Using Technology by Group***

	N	Mean	SD	Min.	Max.
RS	19	27:13	16:31	4:08	62:32
TMM	19	39:10	17:00	12:44	65:26

	N	Mean	SD	Min.	Max.
MEM	21	36:56	19:00	12:13	82:00
WOW	19	20:13	12:43	4:00	60:58
Total	78	31:23	17:35	4:00	82:00

Still, one-way ANOVA tests on hours participants used software in the study evidenced significant differences among means at the $p < 0.05$ level [$F(1, 76) = 11.43$, $p = 0.001$)]. Looking more closely, Tukey post hoc tests showed where significant variation occurred: Both TMM (+19 hours, $p = 0.002$) and MEM (+16 hours, $p = 0.008$) used the technology significantly more than WOW. In spite of that, no group's mean met the targeted goal of 45 hours total, 3 hours per week for the 15 total weeks that made up the duration of the study.

Finally, Welch ANOVA tests indicated significant difference between the means of the total number of hours that female participants used the technology (M = 36:30, SD = 19:36) and the means of the total number of hours male participants used the technology (M = 24:14, SD = 12:30), with significance at the $p < 0.05$ level of ($p = 0.001$).

Correlation Between Hours Using Technology and Posttest Scores

Pearson product-moment correlation tests measured if and to what level of significance the amount of time learners spent using language-learning software related to posttest scores. Results showed no significant correlation overall between total hours using language-learning software and reading, writing, listening, speaking, grammar, or posttest totals, indicating that the increase in overall software use did not correlate with increased scores.

In the RS, TMM, MEM, and WOW groups individually, Pearson product-moment correlation tests also showed that total hours using the software did not correlate with posttest scores in any section of the test or test totals.

Summary of Findings

The WOW group showed significant improvement in three of the five skills (i.e., writing, listening, and speaking) and significant improvement overall. The second most successful software appeared to be MEM, whose group showed statistically highly significant improvement in writing and significant improvement overall. Finally, the RS and the TMM groups showed statistically significant improvement in writing but not overall.

Table 4. ***Summary of Findings***

Group	M Hours	Reading	Writing	Listening	Speaking	Grammar	TOTAL
RS	27:13	—	$p < .001$	—	—	—	—

Group	M Hours	Reading	Writing	Listening	Speaking	Grammar	TOTAL
TMM	39:10	—	*p* = .002	—	—	—	—
MEM	36:56	—	*p* < .001	—	*p* = .001	—	*p* = .001
WOW	20:13	—	*p* = .017	*p* = .05*	*p* = .025	—	*p* = .002

* = Wilcoxon's test showed nominally insignificant improvement (p = 0.051).

Discussion

The researcher investigated the following research question: How do Rosetta Stone, Tell Me More, Memrise, and ESL WOW impact the English performance on a high-stakes standardized test meant to measure reading, writing, listening, speaking, and grammar?

Although RS and TMM groups both improved only on writing, and although the RS group's writing improvement proved statistically highly significant, the RS group's starting out at a lower level seems to have offered more room for fundamental improvement. Still, at least among this group of learners, RS failed to motivate participants to approach the target of 3 hours per week, 45 hours total over the 15-week duration of the study. This finding seems in line with Nielson's (2011) findings, which showed participants using RS much less than anticipated and receiving little overall benefit. These findings also seem consistent with previous criticisms of RS. For instance, Lopez-Lopez (2012) argued that RS lacked cognitive challenge, offered weak connections between vocabulary, grammar, and the learners' lives, did not allow for engaging collaboration, offered no real reading or writing practice across the curriculum, and ignored opportunities for learners to build reading or writing strategies. The RS group's relatively infrequent use of the software might have resulted from a perception that the program lacked connections to learners' lives; indeed, Zheng (2012) found that Chinese learners of English lost interest in areas of English learning not emphasized in class. The findings of this study also seem consistent with DeWaard's (2013) evaluation that, in spite of its self-marketing as a one-stop solution for language learning, RS fails to represent "a viable option for language learning" (p. 61). In contrast, the amount of time TMM learners used the software seems inconsistent with Nielson's (2011) findings, in which participants mostly failed to use TMM according to protocol.

It seemed that TMM and MEM garnered the most attention from participants. Perhaps TMM's optional focus on business or everyday English, as well as the level of online instruction matching participants' scores on TMM placement exams upon initiation of the program, offered more interest to learners and offered more meaningful, comprehensible engagement. TMM's having face validity to learners, and in that way earning more attention, would be consistent with Hashim and Yunus's (2010) survey data, in which learners of English reported thinking TMM was easy to use, helpful for learning English, and generally suitable for language learning. In addition, TMM group's improving only on writing might reflect findings from Zhang and Guo (2012), who concluded that as Chinese learners' proficiency increased, their motivation decreased, especially intrinsic

motivation. The TMM group began at the highest performance level on pretests, which might help explain an overall lack of significant improvement.

The MEM group improved the most in productive skills (writing and speaking). Previous studies have pointed out Chinese learners of English at the university level improved vocabulary levels most markedly during their first and second years but then plateaued during their third (Cui & Wang, 2006; Tan, 2006). Others have accounted for this stabilization by indicating pedagogical shortcomings, such as few opportunities to use rich vocabulary in classrooms (Laufer, 1998). The findings here indicate that sustained vocabulary instruction using out-of-class activities with MEM may impact both writing and speaking test performance among Chinese learners of university level. Regarding the group's relatively frequent use of the software, it may be that learners' ability to view individual achievement in relation to peers and to other learners worldwide on leaderboards sparked interest through competition.

As mentioned above, the WOW group showed significant improvement in writing, listening, and speaking, as well as overall. Although the WOW group made perhaps the most comprehensive improvement, learners in this group also used the software the least amount of time on average. This might reflect findings from Zhang and Guo (2012), which indicated that as Chinese learners' proficiency of English increased, their motivation decreased, especially intrinsic motivation. The fewer amount of hours using the software, then, might indicate lower motivation, which has correlated in the past with higher proficiency among similar learners.

Interestingly, findings in this study did not seem to confirm those of Peng (2014), who examined Chinese students' perceptions of which skills posed the greatest challenges and who reported that Chinese learners of English believed speaking was the hardest skill to learn, with listening seeming to be the easiest. In this study, RS and TMM (which offered the most speaking opportunities of all the software) did not seem to relate to improved speaking test performance.

Finally, female participants showed significant difference from male participants in three main areas: pretest scores, posttest scores, and average number of hours using software out of class during the study. These findings support earlier research indicating more motivated female English language learners in China (Lamb, 2004; Liu, 2009, 2012; Yang, Liu, & Wu, 2010).

Limitations

Gass and Mackey (2012) identified limitations inherent in (quasi-)experimental classroom-based research, naming intervening variables such as frequency and intensity of English engagement learners experienced outside of the study. The quasi-experimental design used in this study also opened itself up to sampling limitations. Because convenience sampling makes use of participants who are both available and willing to be studied, the groups may not represent the population (Creswell, 2012). Small samples sizes, as well as an RS group with a greater number of males to females, also limited how reliably conclusions can follow data analysis or comparisons be made between the groups.

Another limitation concerned the infrastructure at the research site. Some learners reported computer crashes or slow Internet, especially when trying to use RS and TMM.

This situation is not new in China. Zou (2011) explored CALL use among Chinese learners and found that differences in computer-based facilities (with fixed, teacher-centered setups or movable seats) posed challenges to implementing CALL uniformly.

Implications and Future Research

With English test performance remaining important to Chinese ELLs both in China and elsewhere, teachers and researchers need to continue to explore motivating and engaging activities. Zheng (2012) warned that what a teacher emphasizes may shape a learner's opinion of what aspects of English deserve attention. This study indicated that text- and vocabulary-centered MEM and WOW might impact writing and speaking test performance more than RS and TMM, which promote themselves as one-stop solutions to language learning. MEM and WOW remain free, open-source tools available to any learner in the world with an Internet connection and therefore might represent the most potent supplemental tools for settings with limited resources. Future research could build on the findings here by using mixed-methods designs to paint more all-inclusive pictures of the performance and perceptions of learners using this study's software, such as explanatory-sequential or multiphase designs.

About the Author

Justin Nicholes has taught ESL and EFL for twelve years, including seven years in China. Previously published in Language Education in Asia, he will soon begin PhD studies at Indiana University of Pennsylvania.

References

Benson, P. (2007). Autonomy in language teaching and learning. *Language Teaching, 40*, 21-40.

Brynko, B. (2008). Auralog: Speaking your language. *Information Today, 25*(10), 41.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Upper Saddle River, NJ: Pearson Education.

Cummins, J., Brown, K., & Sayers, D. (2007). *Literacy, technology, and diversity: Teaching for success in changing times*. Boston, MA: Allyn & Bacon/Pearson.

Cui, Y., & Wang, T. (2006). A study on the development paths and interrelations on receptive vocabulary size, productive vocabulary size, and vocabulary depth knowledge. *Modern Foreign Languages, 29*, 392-400.

Demir, S., & Korkmaz, G. (2013). The effectiveness of foreign language learning software on students' listening and speaking skills: A case of Rosetta Stone. *Electronic Journal of Social Sciences, 12*(45), 35-51.

DeWaard, L. (2013). Is Rosetta Stone a viable option for second-language learning? *ADFL Bulletin, 42*(2), 61-72. doi:10.1632/adfl.42.2.61

Gass, S. M., & Mackey, A. (2012). *Data elicitation for second and foreign language research*. Beijing, China: Foreign Language Teaching and Research Press.

Hartog, J., Sun, Y., & Ding, X. (2010). University rank and bachelor's labour market positions in China. *Economics of Education Review, 29*, 971-979. doi:10.1016/j.econedurev.2010.06.003

Hashim, H., & Yunus, M. (2010). Learning via ICT: 'TELL ME MORE.' *International Journal of Learning, 17*(3), 211-223.

Holec, H. (1981). *Autonomy in foreign language learning*. Oxford, United Kingdom: Pergamon. Institute of International Education. (2013). Open doors data: International students, leading places of origin. Retrieved from iie.org/Research-and-Publications/Open-Doors/Data/International-Students/Leading-Places-of-Origin/2011-13

Lamb, M. (2004). Integrative motivation in a globalizing world. *System, 32*, 3-19.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19*(2), 255-271.

Li, F. L., Zhao, Y. D., & Tian, Y. P. (2010). Job search and over-education: Evidence from China's labour market for postgraduates. *Perspectives in Education, 28(*2), 41-50. Retrieved from perspectives-in-education.com

Li, J. (2007). An attempted evaluation of computer assisted language learning in China. *Canadian Social Sciences, 3*(3), 109-113. Retrieved from cscanada.net/

Li, Q. (2014). Differences in the motivation of Chinese learners of English in a foreign and second language context. *System, 42*, 451-461. doi:10.1016/j.system.2014.01.011

Liao, X. (2004). The need for Communicative Language Teaching in China. *ELT Journal, 58*(3), 270-273. doi:10.1093/elt/58.3.270

Liu, M. (2012). Motivation in Chinese university EFL learners in varying learning contexts. *TESL Reporter, 45*(2), 17-39.

Liu, M. (2009). *Reticence and anxiety in oral English lessons*. Berne, Switzerland: Peter Lang International Academic Publishers.

Liu, M. (2006). Anxiety in Chinese EFL students at different proficiency levels. *System, 34*(3), 301-316.

Liu, M., & Jackson, J. (2011). Reticence in Chinese EFL students at varied proficiency levels. *TESL Canada Journal, 26*(2), 65-81. Retrieved from teslcanadajournal.ca/

Lopez-Lopez, S. (2012). Rosetta Stone TOTALe® K-12: Spanish, Latin America. *TESL-EJ, 16*(1), 1-9.

Lu, Z., Li, P., & Du, P. (2009). Interactive patterns in an English audio-video speaking class in CALL environments. *International Journal of Pedagogies, 5*(2), 49-66. doi:10.5172/ijpl.5.2.49

Ma, R. (2012). Vocabulary proficiency instruction for Chinese EFL learners. *Theory & Practice in Language Studies, 2*(6), 1199-1205. doi:10.4304/tpls.2.6.1199-1205

Nielson, K. B. (2011). Self-Study with language learning software in the workplace: What happens? *Language Learning & Technology, 15*(3), 110-129.

Ning, H., & Hornby, G. (2013). The impact of cooperative learning on tertiary EFL learners' motivation. *Educational Review, 66*(1), 108-124.

Nisbet, D. L., Tindall, E. R., & Arroyo, A. A. (2005). Language learning strategies and English proficiency of Chinese university students. *Foreign Language Annals, 38*(1), 100-107.

Niu-Cooper, R. (2012). Unexpected realities: Lessons from China's new English textbook implementation. *International Journal of Education Policy and Leadership, 7*(2), 1-17. Retrieved from journals.sfu.ca/ijepl/

Orlik, T. (2012, December 8). Chinese survey finds a higher jobless rate. The Wall Street Journal. Retrieved from online.wsj.com/

Peng, S. (2014). Analysis of perceived difficulty rank of English skills of college students in China. *Canadian Social Science, 10*(5). doi:10.3968%2F4832

Peng, J., & Woodrow, L. (2010). Willingness to communicate in English: A model in the Chinese EFL classroom context. *Language Learning, 60*(4), 834-876.

Qi, Y. (2011). Improving graduates' employment competitiveness: A practice of Peking university. *Education for Information, 28*, 101-113. Retrieved from iospress.nl/journal/education-for-information/

Ruan, Q. (2008). A survey on CALL study in China in recent decade. *Canadian Social Science, 4*(1), 41-44.

Shen, L., & Suwanthep, J. (2011). E-learning constructive role plays for EFL learners in China's tertiary education. *Asian EFL Journal, CEBU Issue, 54*. 4-29.

Shield, L. (2009). CALL: Using what we know to avoid reinventing the wheel. *Indian Journal of Applied Linguistics, 35*(1), 11-24.

Stoltzfus, A. (1997). *The learning theory behind the Rosetta Stone Language Library from Fairfield Language Technologies*. Paper presented at National Association for Bilingual Education Annual Meeting. Albuquerque, NM.

Tan, X. (2006). A study on Chinese EFL learners' productive vocabulary development. *Foreign Language Teaching and Research, 38*(3), 202-207.

Wang, N. X., & Morgan, W. J. (2009). Student motivations, quality and status in adult higher education (AHE) in China. *International Journal of Lifelong Education, 26*(4), 473-491. doi:10.1080/02601370903031314

Yang, L., Liu, M., & Wu, W. (2010). An investigation of Chinese undergraduate non-English majors English learning motivation. In Z. Lu, W. Zhang, & P. Adams (Eds.), *ELT at tertiary levels in Asian contexts: Issues and researchers* (pp. 48-62). Hong Kong, China: Hong Kong Polytechnic University.

You, X. (2004). The choice made from no choice: English writing instruction in a Chinese university. *Journal of Second Language Writing, 13*(2), 97-110. doi:10.1016/j.jslw.2003.11.001

Zhang, Y., & Guo, H. (2012). A study of English writing and domain-specific motivation and self-efficacy of Chinese EFL learners. *Journal of Pan-Pacific Association of Applied Linguistics, 16*(2), 101-121.

Zhao, G. Z. (2009). Higher education scale and employment relationship in China. *US-China Education Review, 6*(5), 16-20. Retrieved fromdavidpublishing.com/journals

Zheng, Y. (2012). Exploring long-term productive vocabulary development in an EFL context: The role of motivation. *System, 40*(1), 104-119. doi:10.1016/j.system.2012.01.007

Zhou, M. M., & Xu, Y. (2012). A self-determination approach to understanding Chinese university students' choice of academic majors.*Individual Differences Research, 10*(1), 49-59. Retrieved from idr-journal.com

Zou, X. (2011). What happens in different contexts and how to do learner autonomy better? *Teacher Development, 15*(4), 421-433.