Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2022, Volume 26, Issue 1
pp. 1–30

ARTICLE

Investigating learner autonomy and vocabulary learning efficiency with MALL

Nigel P. Daly, *National Taiwan Normal University*

Abstract

The road to second language competence is a long and arduous one, and much of its effort involves learning to recognize and use vocabulary. Fortunately, anytime-anywhere learning with smart phones and smart apps offer a means to lessen the burden and make vocabulary learning more efficient. Accordingly, this study investigated 134 students across four months and evaluated the effectiveness of their individual vocabulary learning strategies (only flashcard app; paper-based notes and wordlists; both notes and flashcard app) in terms of three different vocabulary test scores. The Kruskal-Wallis rank sum test and pairwise comparisons revealed that the Only App Group had significantly higher test scores than both the Only Notes Group and the blended Notes/App Group with medium and small effect sizes (r = 0.49 and 0.27, respectively). A Fixed Effects model was run to determine the extent study strategies in addition to gender, (TOEIC) proficiency, time spent studying, time spent using the app, and frequency of studying, were correlated with test scores. In this moderator analysis, the Only App Group strategy was no longer statistically significant and was replaced by the factor "total time using the app" (p = .005) which was positively correlated with test scores.

Keywords: *Mobile Assisted Language Learning (MALL), Second Language Vocabulary Acquisition, Intentional Learning, Autonomous Learning*

Language(s) Learned in This Study: *English*

APA Citation: Daly, N. P. (2022). Investigating learner autonomy and vocabulary learning efficiency with MALL. *Language Learning & Technology*, *26*(1), 1–30. https://doi.org/10125/73469

Introduction

The sophistication and ubiquity of computer assisted language learning (CALL) and mobile assisted language learning (MALL) educational technologies can offer the language learner useful new affordances to assist vocabulary learning in interesting and effective ways. The need for efficient learning has perhaps never been felt so acutely, especially since language learners have been given new benchmarks in vocabulary learning, with earlier proficiency targets of 3000–4000 word families being superseded by new proposed benchmarks of up to 8000–9000 word families (Nation, 2006; Schmitt & Schmitt, 2014). In any learning context, but especially in the foreign language one with no natural environmental input or opportunities for language practice, the task of learning large amounts of vocabulary in short periods of time is a necessary component of language development and depends largely on individual effort and brute memorization. Fortunately, MALL technologies and applications can serve as a more interesting and potentially more efficient alternative to mind-numbing work with wordlists and rote copying out of words, definitions, and sentences. But how—in combination with more traditional means of using paper wordlists and writing out words, or is it enough to only use a MALL app?

MALL Meta-analyses

With cheaper prices for both hardware and mobile phone / internet services, people around the world are turning to mobile phones as their primary means of telecommunication as well as access to important resources for learning. According to the NewZoo global mobile market report dated May 2018, around 25

countries in the world had at least two thirds of its population using smartphones, with Taiwan ranking 5th with 72% smartphone usage. For this reason, there has been a recent rise in the number of *mobile assisted learning* (MAL) studies as well as review articles (e.g., Burston, 2015; Elaish et al., 2017; Frohberg et al., 2009; Klimova, 2018; Liu et al., 2014; Sung et al., 2015; Sung et al., 2016); Sung et al. (2016)'s meta-analysis on 110 MAL articles over twenty years calculated a medium mean effect size of 0.523 (Hedge's *g*; $r \cong 0.31$). As for studies focusing on mobile assisted language learning, Elaish et al. (2017) indicated that approximately 80% of MALL studies show the effectiveness of using mobile devices for language learning (e.g., Agca & Ozdemir, 2013; Amer, 2010). Sung et al. (2015) and Taj et al. (2016) performed meta-analyses on the effects of mobile devices on language learning and found, similar to Sung et al. (2016), mean effect sizes of 0.55 and 0.425, respectively. These may be interpreted as either *medium* effect sizes (Cohen, 1988) or $small_{PO}$ if Plonsky and Oswald (2014)'s revised estimates, specifically for L2 research, are considered (see Appendix A and the Methods section for further discussion on effect size interpretation and approximate conversions; this study will use Plonsky and Oswald's more conservative interpretation for *r* using $small_{PO}$ or $medium_{PO}$ as designations of their interpretations).

Meta-analyses on MALL for Vocabulary Acquisition, the Paucity of Effect Size Data, and Benchmarks

In their review of studies on mobile English language learning, Elaish et al. (2017) pointed out that the most researched topic in MALL studies from 2010 to 2015 was vocabulary acquisition, accounting for 26 out of their sample of 133 articles, or 19.5%. For this reason, there have been a number of review articles on MALL for vocabulary learning, such as Afzali et al. (2017) and Çelik and Yavuz (2018). Although not all studies on MALL and vocabulary acquisition showed statistically significant learning outcomes (e.g., Fisher et al., 2009; Song & Fox 2008; Stockwell, 2007, 2010; Tosuna, 2015), most of these studies predated large, touch screen mobile phones with internet ubiquity that are commonplace today. The two recent meta-analyses of Mahdi (2018) and Lin and Lin (2019) have revealed that the use of mobile devices for vocabulary learning have $medium_{PO}$ to $large_{PO}$ effect sizes. These two metanalyses on vocabulary acquisition are taken as benchmarks due to their rigorous inclusion criteria: Mahdi (2018) retained a total of 33 out of 83 studies and Lin and Lin (2019) only 16 out of 219 studies, with both studies taking as a major consideration the reporting of statistics sufficient to calculate effect sizes. Burston's (2015) MALL meta-analysis did not even report effect sizes since only half of the final selected 19 studies (out of an initial 291) listed such statistics. Mahdi (2018) examined 16 studies with 986 participants (of which 15 postdated 2010, the year the iPad was released) and found that mobile devices had $medium_{PO}$ effect sizes for vocabulary learning in general (Hedge's $g = 0.67$ [$r \cong 0.4$]), as well as specifically for receptive ($g = 0.62$ [$r \cong 0.4$]; 12 studies) and productive vocabulary learning ($g = 0.61$ [$r \cong 0.4$]; 4 studies). Lin and Lin (2019) meta-analyzed 33 studies published between 2015 and 2018 and calculated a $large_{PO}$ effect size ($g = 1.0$, range: 0.05–4.7); relevant to this study, they also conducted a moderator analysis and found effect sizes for MALL use (a) in informal out-of-classroom settings were $large_{PO}$ ($g = 1.02$), (b) $medium_{PO}$ ($g = 0.6$) for a longer duration of more than 10 weeks, and (c) for full learner autonomy also $medium_{PO}$ ($g = 0.77$).

Vocabulary Learning Strategies

If becoming a proficient second language user requires the learning of 8000 to 9000 word-families in formal educational settings, the challenge is to efficiently learn and memorize a large amount of vocabulary in the short amount of time available to second language learners. To assist learners in this endeavor, several books (e.g., Nation, 1990, 2001, 2008; Pavičić Takač, 2008) have thus been dedicated to *vocabulary learning strategies* (VLS), such as memorization involving wordlists, oral rehearsals, associations and mnemonics, copying words on paper, and making and using flashcards. In fact, researchers have identified 58–69 different VLS (e.g., Jiménez Catalán, 2003; Pavičić Takač, 2008; Schmitt, 1997), and have shown that frequent VLS usage positively correlates with vocabulary size (Alahmadia et al., 2018). Schmitt's (1997) influential taxonomy of VLS was a synthesis of previous research (e.g., Cook & Mayer, 1983; Nation, 1990; Oxford, 1990), which divided VLS into two main categories, discovery and consolidation. The former involves the ways learners discover the meaning of words (e.g., using dictionaries or asking the

teacher or classmates for the meaning), while the latter relates to how the learner consolidates and remembers the word meaning after it is discovered. Directly relating to the present study are the consolidation strategies, which Schmitt (1997) further sub-divided into *memory*, *cognitive*, and *metacognitive*. Examples of these strategies can be found in Jiménez Catalán's (2003) study of 580 Spanish EFL learners. In this work, she found the commonly used *memory* strategies included studying by using cognates, saying the new word out loud, and forming an image of the word's meaning. For *cognitive* strategies, learners were taking notes in class, using the vocabulary section in the textbook, and employing verbal repetition. Finally, for *metacognitive* strategies, they were using English media, testing oneself with word tests, and continuing to study the word over time (pp. 76–77).

Paper and Digital Flashcards

It is interesting to note that out of the nine cognitive strategies in Jiménez Catalán's (2003) 580-respondent study, paper flashcards (PFs) ranked in 8th place with only about 7% of the respondents using them. However, the apparent learner apathy to PFs has not been shared by researchers, who have investigated and extolled their effectiveness (e.g., Fitzpatrick et al., 2008;-Nation, 2001, 2008; Nation & Waring, 1997). This is most likely due to their user-friendliness and consistency of use among users that facilitate research comparability of control and treatment groups. Ease of creation and use is also why several digital flashcard (DF) apps have emerged in recent years, such as Quizlet, Cram, Anki, and others. However, unlike their paper-based predecessors, DFs have been embraced and appear to enjoy widespread use with the number of Quizlet users exceeding 50 million a month (Quizlet, 2022). These apps do not require the labor-intensive creation and inconvenience of portability and accessibility of paper flashcards.

In light of this, several studies have investigated the design of DFs (e.g., Anaraki, 2009; Ou-Yang & Wu, 2017) and their empirical or perceived effectiveness (e.g., Anaraki, 2009; Davie & Hilber, 2015; Nikoopour & Kazemi, 2014) by comparing the effects of learning vocabulary with DFs versus PFs, often employing pre- and post-test study designs. These studies show that DFs are typically more effective (e.g., Azabdaftari & Mozaheb, 2012; Başoğlu & Akdemir, 2010; Kiliçkaya & Krajka, 2010; Nikoopour & Kazemi, 2014). The relative effectiveness of DFs is most likely due to the anytime, anywhere convenience of having hundreds or thousands of flashcards in the cell phone in your pocket. Keeping this in mind, it seems odd that researchers are still comparing paper versus digital flashcard use. And while a recent study by Dizon and Tang (2017) comparing the effectiveness of DFs and PFs found a non-significant difference between the two methods, the study results are questionable from a number of perspectives: small sample sizes; pre-test and post-test content consisted of randomly selected words from the most common 1000- to 2000-level words, most of which should have already been known by the university learners; and most importantly, the PF Group was trained in three different vocabulary learning strategies, while the DF Group was not. The latter variable was added to level the playing field, but if that is the case, one may legitimately ask—especially in areas where smartphone ownership is virtually 100% among the learning population—why are PFs still being investigated at all?

Under-researched Areas

Given the positive findings of MALL research as well as the above-mentioned benefits of MALL technologies for vocabulary learning, this study was motivated by the desire to shed light on three under-researched issues:

1) How MALL DF technologies are used by learners autonomously outside the classroom;
2) How these technologies are used longitudinally;
3) Whether a blended MALL-related strategy is more effective than an only-MALL strategy, as measured by test performances.

Regarding the first issue, the research literature on technology-mediated second language vocabulary development suffers from a lack of information on learner autonomy in terms of learners' "capacity for independent learning" and how vocabulary develops outside the controls and constraints of the classroom or research laboratory (Elgort, 2018, pp. 18–19). As for the second issue, longitudinal studies in MALL for

vocabulary learning have been sparse: in Lin and Lin's (2019) and Burston's (2015) MALL meta-analyses, only 7% and 3% of their respective studies were over 10 weeks in duration. In terms of the third under-researched issue, although there have been MALL studies comparing different vocabulary studying strategies, Elgort (2018, pp. 19) has lamented that they typically compare technology versus non-technology mediated treatments (e.g., DFs and PFs), and fail to take into account other possible strategies, such as blended strategies using a combination of paper wordlists and DFs, which would probably be a natural choice for many students if both tools were presented to them.

In summary, the previous studies have tended to focus on the controlled use of DFs in short duration studies. While they have demonstrated the positive effects of MALL learning in general and specifically DFs for vocabulary learning, the literature continues to remain silent on alternative strategies such as combining paper wordlists with mobile technologies, like DFs. This combined strategy is a natural choice for many learners for a few reasons: paper wordlists are often easier to create and obtain compared to PFs; they are easier and faster to scan through; and they are tangible, which is reassuring for many learners. As such, three *general* consolidation strategies were investigated in this study: (a) only paper wordlists and no DF app use, (b) the combined use of wordlists and DF app use, and (c) only using the DF app (as accessed by either smart phone or computer). Because the focus of this study is on wordlists and flashcards, other more *specific* strategies were not investigated, such as *cognitive* strategies like verbal repetition and written repetition, and *memory* strategies like saying the new word when studying, forming an image of the word's meaning, studying the spelling, and using cognates when studying.

The DF app chosen for this study was Quizlet (hereafter referred to as "app") because it is free, user-friendly, very popular, consistently praised in online language learning app reviews, and has won awards, such as being recently voted "Best Free Language-Learning App for Rote Memorization" by PC Magazine (Duffy, 2020). Several studies (Chien, 2015; Dizon, 2016; Dizon & Tang, 2017; Jackson, 2015) have already noted students' appreciation of and preference for Quizlet over other DF apps, such as Study Stack or Cram.

Study Aims and Hypothesis

This study is exploratory in nature, with two main aims: (a) to observe learner autonomy over 4 months in terms of the extent of the app's habitual use (or non-use) outside of the classroom and how it was used (by means of post-quiz surveys); and (b) to investigate learning efficiency by comparing the effectiveness of three general vocabulary learning strategies: notes/wordlist only, blended wordlist and app use, and only app use. This study also seeks to determine if there were other moderating factors affecting the performance on summative vocabulary tests (three quizzes across 4 months) and to measure the size of their effects. Considering previous meta-analytic research on MALL for vocabulary learning (Mahdi, 2018; Lin & Lin, 2019), the strength of the relationship between the app use and test scores is hypothesized to be of $medium_{PO}$ effect size (r = 0.4–0.6).

Methods

Research Design

The design of this study was non-experimental, insofar as the sample population was not randomized and the independent variable of studying strategy was not controlled; however, it still explored causal relationships ex post facto in a causal-comparative research design (Fraenkel & Wallen, 2006). These two factors were not manipulated because the dependent variable of test scores largely contributed to the students' final course scores and it was thought to be fairer to let them choose their own studying method, which would also allow for a better understanding of the autonomous and authentic use of the flash card app as naturally adopted habits and preferences. After each of the three quizzes, the students used their smartphones to fill out a digital survey on a Google Form that asked them about their personal characteristics (e.g., sex, TOEIC score) and study habits for that particular quiz. Both the quiz scores and information from this form were analyzed to compare the variation between groups (non-parametric tests) and within groups (Fixed Effects Model in a panel regression) to see how the three study strategies as well

as other possible moderators accounted for the variance in the dependent variable of test scores.

Participants

The participants were enrolled in a six-month government sponsored business training program in Taipei, Taiwan. Usually, the students who enroll in this intensive program are in their mid-twenties, have a minimum TOEIC score of 550 and a university degree, and tend to be highly motivated to improve their English ability and learn business skills to seek a career in international business. The study investigated the vocabulary learning strategies and learning performance in a business English course that tried to balance Nation's "four strands" of meaning focused input, meaning focused output, language focused learning, and fluency development (Nation & Yamamoto, 2012). Data from two sessions were collected: 68 students were from the Summer/Fall session of 2016, and 66 others were from the Winter/Spring session of 2018. Both groups of students were exposed to the same vocabulary learning components (i.e., vocabulary, in-class practice and testing) over a four-month period. The use of Quizlet was encouraged, with six ready-made study sets (totaling over 600 words), and frequent quizzes and surveys were given.

Instrument

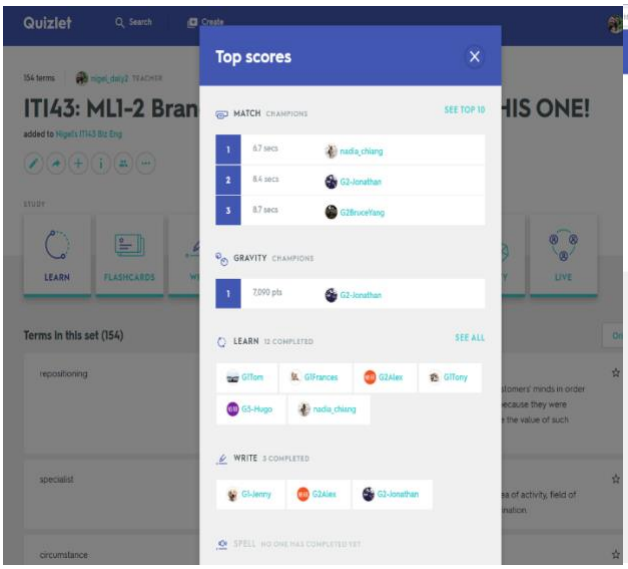
Every three weeks or so, a paper-and-pencil summative quiz of 30 questions was given, which contained questions testing mostly receptive ability (multiple choice or matching) but also some productive ability (fill-in-the-blanks) at a ratio of about 4:1 (see Appendix B for sample questions). The content of the quizzes was cumulative, with about 125–150 new words added for each quiz, such that learners were responsible for 125 words for Quiz 1, and over 500 by Quiz 4. Approximately 60% of the test items came from the most recent set, and the remaining 40% from the previous units. There was a Quizlet study set for each textbook unit containing 60–75 words and followed the theme of the unit. Immediately after these tri-weekly quizzes, a Google Form survey filled out by the students on their smartphones in the classroom (multiple choice and Likert scale items; Appendix C) collected information about their studying habits and use of Quizlet.

From a pedagogical stance, Quizlet is useful for both teachers and students. For teachers, they can sign up for a paid account that enables them to open their own virtual classroom to which their students can join. This is useful because the app's analytics record student activity such as those who complete various study activities or who feature on the top 10 scoreboard for games; this information was used in this study to award extra marks to students. For students, Quizlet offers a wide variety of study activity options to reinforce vocabulary learning and stave off boredom, including basic flashcards, multiple choice review questions (Learn), writing out the word in response to a definition (Write), multimodal dictation where the definition is seen and the word is heard (Speller), test creation options for learners to create their own multiple choice or cloze-style tests (Test), and two speed-based games (Matching, Gravity). See Figure 1 for screenshots of the "Top scores" window, Quizlet interface, "Learn" study activity and "Match" game.

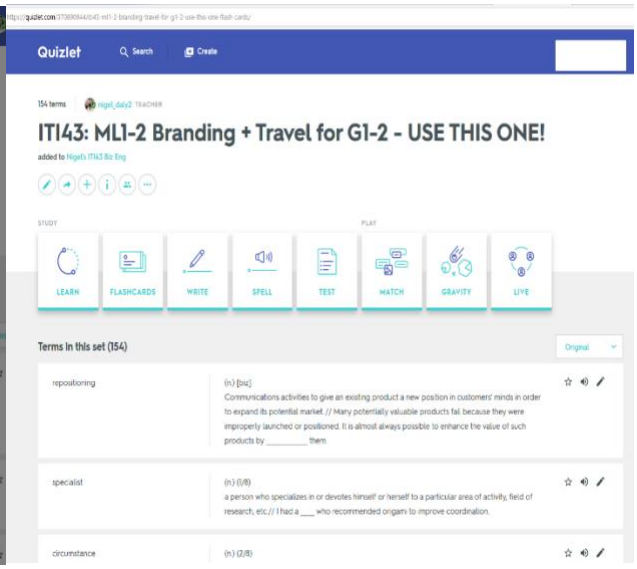
Figure 1

Quizlet Interface (Clockwise from Top Left): a. "Top Scores" Window for Teacher Accounts, b. Study Set Interface with Study Activity Options, c. "Learn" Activity, and d. "Match" Game Activity

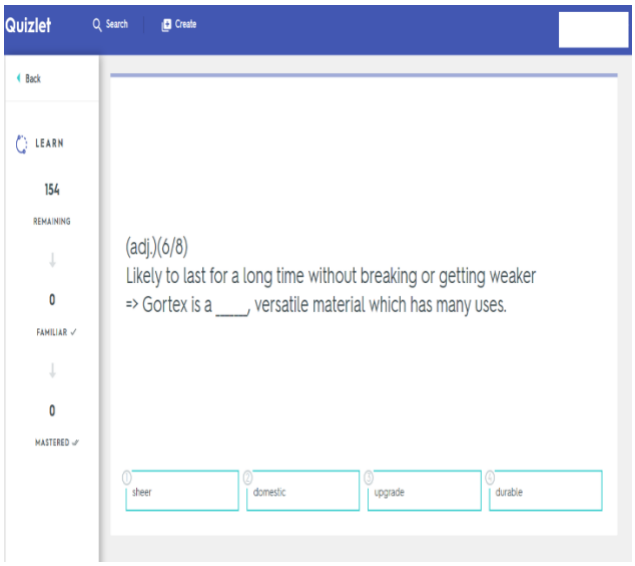
a.



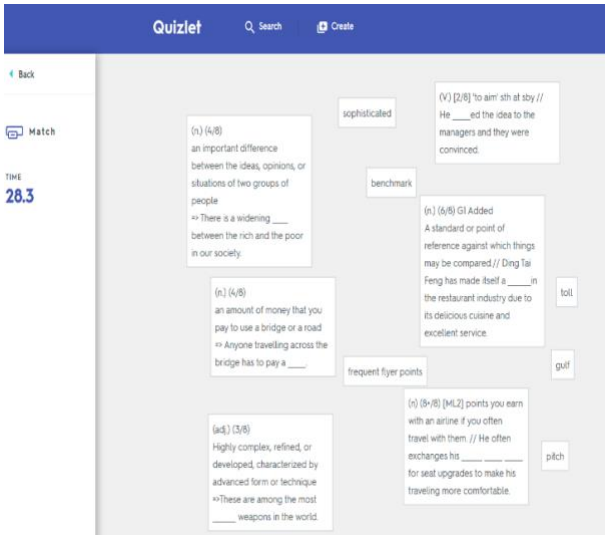
b.



c.



d.



Data Analysis

Nonparametric Tests

The first, second, and fourth quiz scores and surveys for both the 2016 and 2018 sessions were chosen for the data analysis (data from the third quiz was incomplete). The number of valid samples from the combined sessions was 367 (Quiz 1 = 124; Quiz 2 = 125; Quiz 4 = 118). Given that students were allowed to choose their strategy, and therefore not controlled, a repeated measures ANOVA was not possible. As such, a meta-

analytic approach was adopted, with the data from the three quizzes being analyzed as independent events with averages taken and aggregated data combining all three quizzes into the three strategy groups being analyzed and presented in parallel. According to the Shapiro-Wilk Test, none of the test score distributions from any strategy group had a normal distribution, and for this reason, non-parametric tests were applied, and both means and medians were reported in addition to interquartile ranges; neither standard deviations nor confidence intervals were calculable given their parametric assumptions. The statistical programming software *R* (R Core Team, 2020) was used to conduct the statistical analyses.

Although not an experimental study, this causal-correlational investigation requires the use of statistical tests that follow a hypothesis testing logic. Specifically, the null hypothesis states that there is no statistically significant difference among the test-taking confidence rankings of the students who used strategies A, B, or C. The non-parametric Kruskal-Wallis statistic was used to determine whether there was a significant difference (defined as $p < .05$) among the rankings of the three strategy groups from the aggregated data from the three quizzes. Because time was not expected to play a role in strategy use (i.e., improved strategy use over time leading to improved scores) and approximately 45% of the students switched strategies at least once across the three quizzes, it was not treated as an independent variable, and a repeated measures analysis was not conducted. The following assumptions for using the Kruskal-Wallis statistic were met: (a) the independent variable (study strategy) has three or more levels (A, B, and C); (b) the groups defined by the levels of the independent variable comprise different participants; and (c) the dependent variable (test scores) yields rankable data (Turner, 2014, p. 246).

Once the Kruskal-Wallis test suggested a difference of at least two groups, the Wilcoxon Rank Sum Test was applied to each pair to determine any statistical difference between them with alpha set at 0.05; a Benjamini-Hochberg correction was used as the *p-value* adjustment method to control for false discovery rate. For pairwise comparisons with $p < .05$, the Wilcoxon test was used to create a *z* standardized score to calculate the effect size ($r = Z/\sqrt{N}$; Rosenthal, 1994); strategy group effect sizes were then compared to see which group had the largest effect size.

Effect Size Calculations

The effect sizes calculated in this study were interpreted in the context of previous MALL studies as well as Plonsky and Oswald's (2014) revised effect size rules of thumb for L2 research, because they found that Cohen's (1988) original and widely-used guidelines underestimated the effect sizes typically found in L2 research; r_{PO} will be used to designate these *r* value revisions, with r_{PO} = .25, .4, and .6 for small, medium and large effect sizes, respectively. Cohen's *d* or Hedge's *g* could not be calculated due to their parametric assumptions in the calculation of differences in means, so *r* scores as indicating variance explained were used as the effect size measures. To facilitate comparison with previous meta-analyses that used Hedge's *g* for effect size (e.g., Mahdi, 2018; Sung et al., 2016), the *r* values from this study were compared to Cohen's *d* as an approximate proxy for Hedge's *g* from Plonsky and Oswald's (2014) revised *r* and *d* interpretations for L2 research (see Appendix A).

Fixed Effects Models

A second, follow-up analysis used a Fixed Effects (FE) Model to investigate the survey data and confirm the superiority of one strategy or determine whether other, less obvious, learner characteristics correlated more closely with the dependent variable of test scores. The FE model has two requirements (Allison, 2005, p. 2) met by the data in this study: The first is that everyone in the sample has at least two measurements on the same dependent variable (i.e., test scores from three quizzes). The second is that for some of the subjects, the values of the independent variables must be different on at least two measurement conditions (i.e., study strategy, study times, etc.) from one quiz to another. ANOVA was not used because parametric assumptions were not met, such as normal distribution (Figure 4) and homogeneity of variances (Figure 5). At least seven models (of varying combinations of factors) were analyzed using the FE model to determine which combination of factors moderated or accounted for the largest amount of variation among the within-group scores. In addition to study strategy, other moderators considered included quiz, individual students,

device used, location of study events, time spent studying, time spent using Quizlet, and number of separate times spent studying; since the FE model requires at least some within-subject variation in the analyzed variables, gender and TOEIC score were not applicable for analysis. Because strategy switching from one test to another did occur for several students, it may have acted as a moderator in and of itself, so six different possible variations of strategy switching (e.g., only notes to notes/app, only notes to only app, etc.) were included in the FE model to determine whether strategy switching itself was a significant moderator.

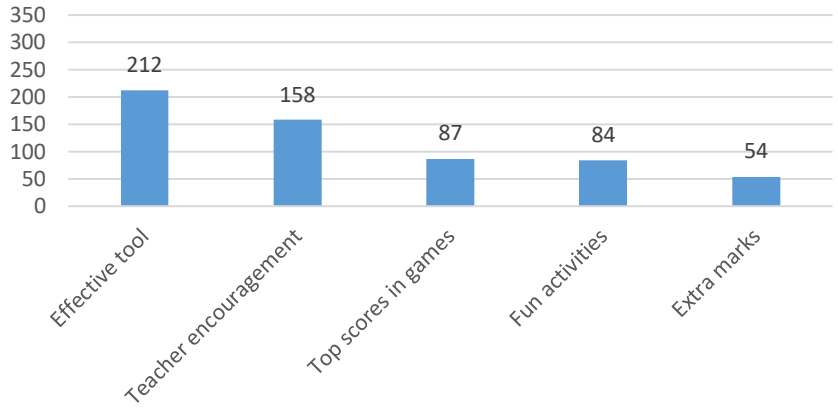
Results

Participant Learning and App Use Motivation

According to the post-quiz surveys, the students were highly motivated to learn insofar as 98% agreed with the survey item that "vocabulary is an important part of language learning" and 94% agreed with item "I have a strong desire to improve my vocabulary ability" (see Appendix C for the survey data). The app was used by most students, and it accounted for most of the studying time across the three quizzes: of the total reported study time (1504.7 hrs), 70.7% (1064.4 hrs) was spent on app use, and the remainder on non-app use. The reasons motivating this widespread use of the app can be seen in the responses to the question asking respondents to choose one or two main reasons for using the app. As Figure 2 illustrates, 61% of the responses indicated their belief that the app was an effective tool, and 45% appreciated the teacher's encouragement; the other reasons were the sense of achievement from completing study set activities and competition with classmates for top scores on the app's games (25%), the fun activities (25%), and finally the extra marks awarded for completing study set activities (16%).

Figure 2

Main Reasons for App Use (N = 342)

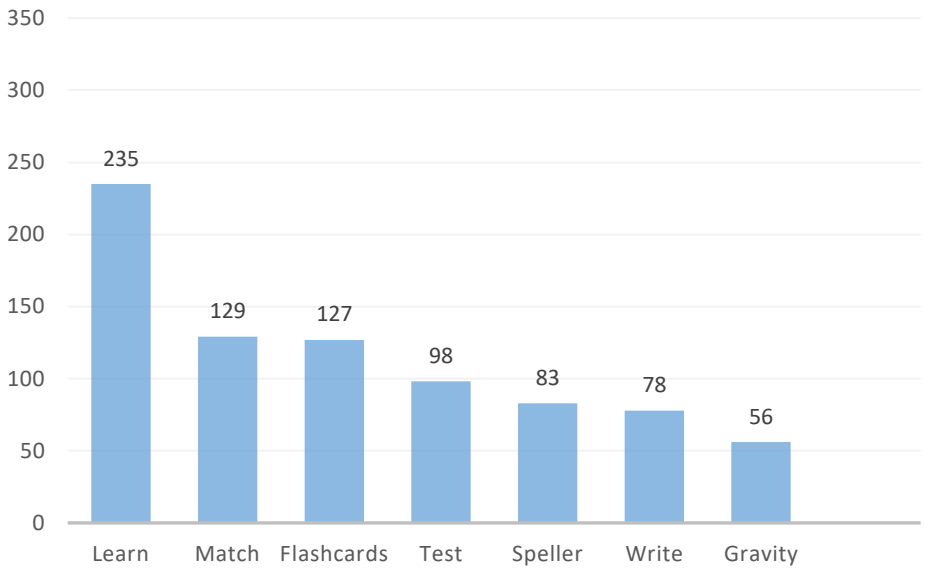


App Usage – Activity and Location

As shown in Figure 3, the most used activity across the three quizzes was the Learn function, which presents the word with multiple choice answers randomly culled from the other digital flashcards (see Appendix C for survey data). The second most used function was the Match game, which required the smartphone users to match a 4x4 grid of word-cards with definition-cards in a race against the clock, with the fastest players' names appearing on the top scorer leader board.

Figure 3

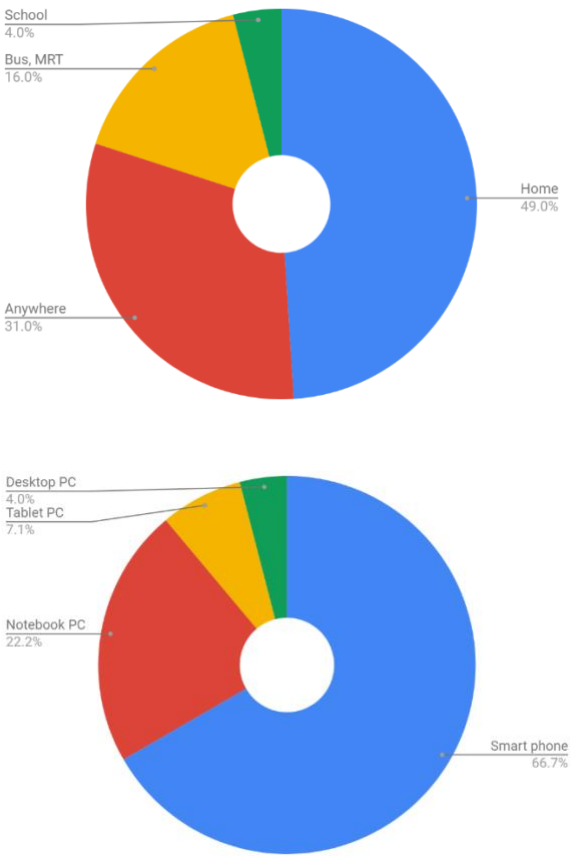
App Study Activity Usage by Learners (N = 342)



When asked which device they primarily used to access the online app, for those that used the app over the three quizzes, the MALL nature of the app became clear. As shown in Figure 4, the app was used over 50% of the time outside the home, and mobile devices accounted for 96% of device usage: smartphones were used by the majority (*n* = 228; 67%), then notebook pcs (*n* = 78; 23%), tablet pcs (*n* = 22; 6%), and finally desktop pcs (*n* =13; 4%).

Figure 4

Main Location and Device for App Use

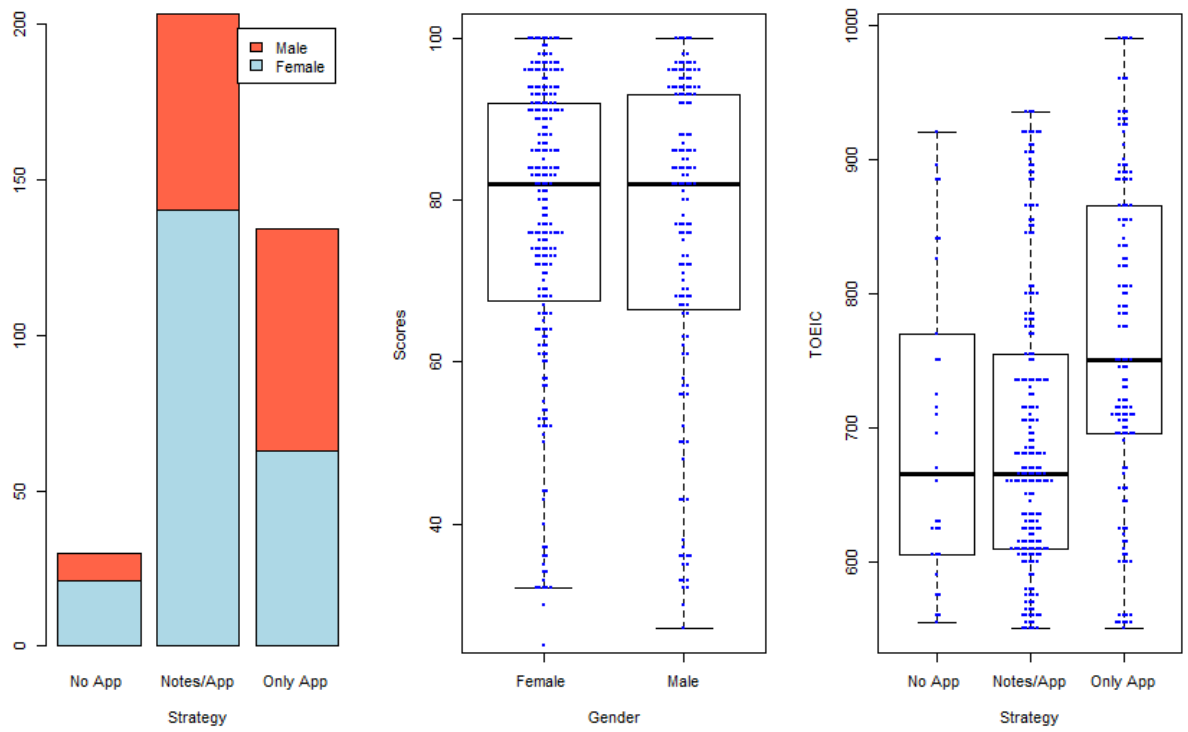


Participants' Gender, Ability, and Score Distributions

There were approximately 122 students per quiz, and across the three quizzes, the gender ratio was about 39% males to 61% females, with the males on average spending a little more time per quiz using the app (3.05 vs. 2.80 hrs). Figure 5 similarly shows that more males preferred the Only App study strategy (53%, or 71 out of 134 of Group C), whereas more females preferred to study using a combination of paper notes and online app (69%, or 139 out of 202 of Group B). However, even though there were some gender differences in strategy preference, there was no difference in medians interquartile score dispersion between males and females (Figure 5). In terms of proficiency, the students had a range of TOEIC scores from 550 to 990 (CEFR B1 to C2), with an average of 720 (high B1) and median of 705. On average, the Only App Group had higher proficiencies with a median score of 750, almost 100 points higher than the other two groups' (660 for No App, and 665 for the Notes/App Group) (Figure 5).

Figure 5

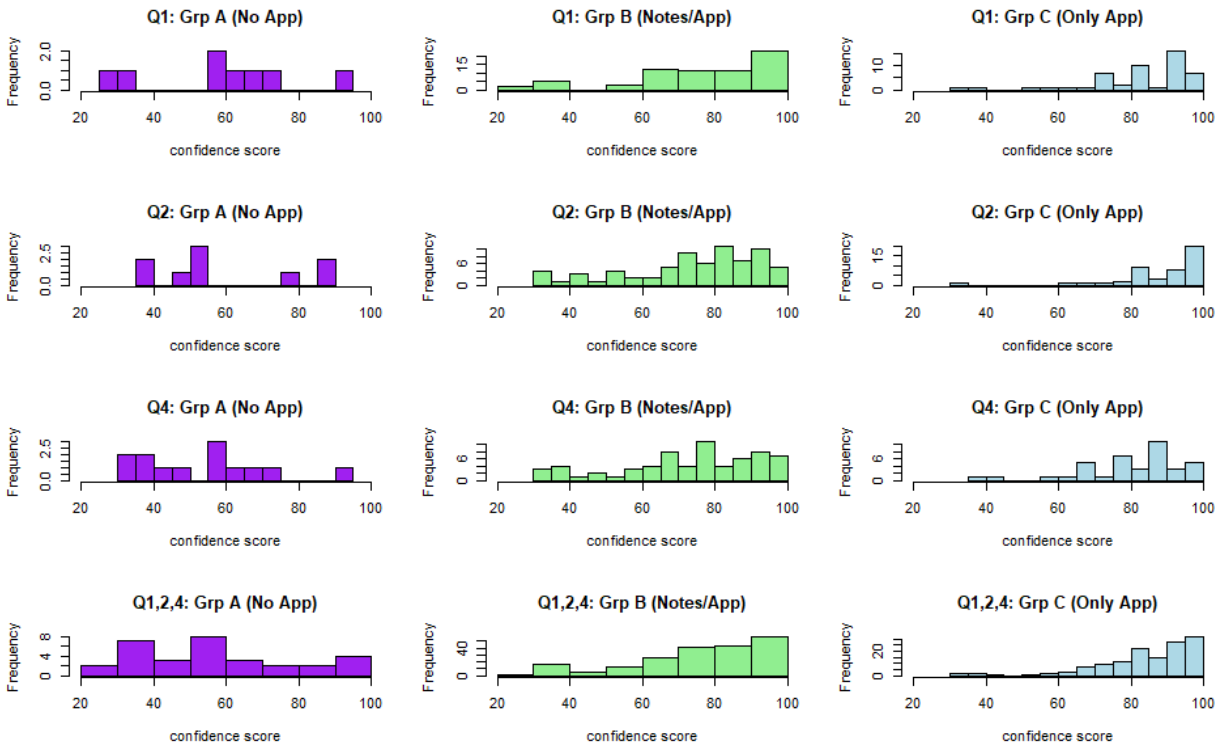
Gender, Strategy and TOEIC Proficiency Correlations



From the nine histograms covering the three strategy groups over three quizzes (Figure 6), two trends are immediately evident. The first is that none of the distributions are normal, with both the Notes/App strategy and Only App strategy being sharply skewed to the right (high test scores). Secondly, the vast majority of students embraced the use of the app (Notes/App and Only App strategies) and very few decided not to (No App).

Figure 6

The Number of Participants Versus Test Scores for the Three Strategies for Each Quiz and Aggregated from all Quizzes



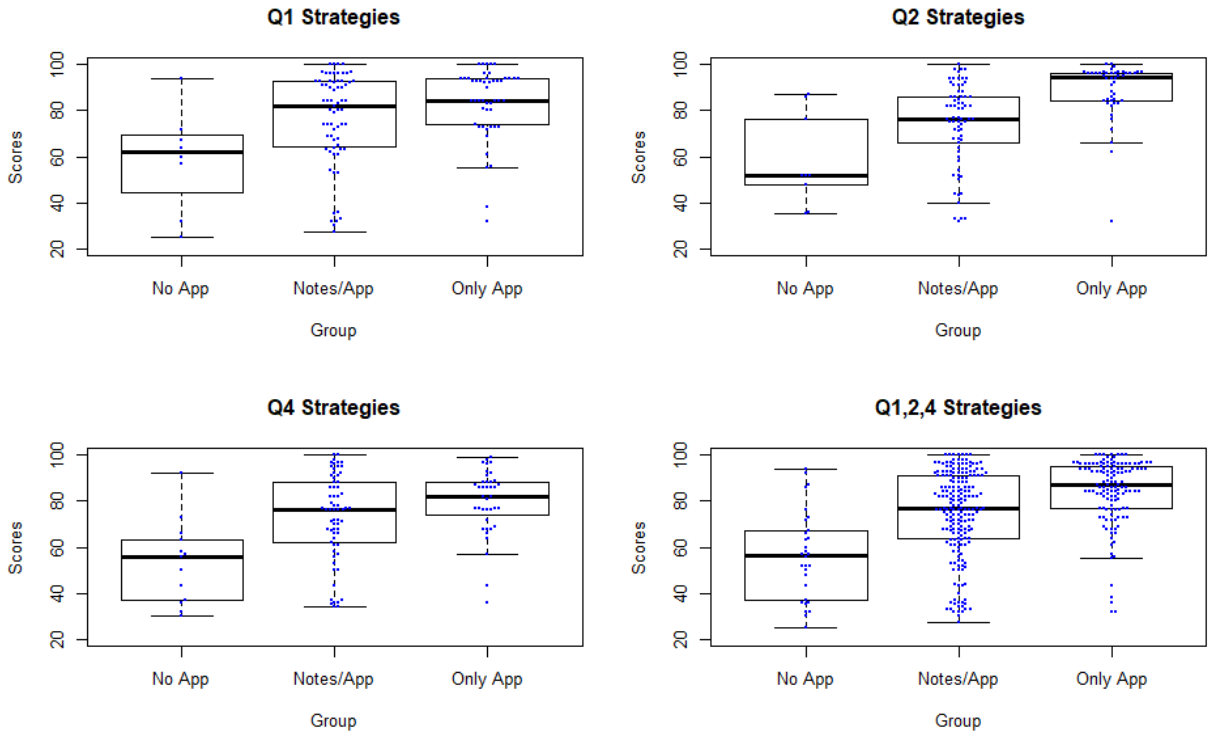
Comparing the Effectiveness of the Study Strategies

In the following, the data from the three quizzes have been presented individually along with their averages and aggregated datasets. The reason for this is to appreciate how sample data can vary from one sampling to the next, which Cumming and Calin-Jageman (2016) have described as the "dance of the means," or in this study's non-parametric case, the dance of the medians. Using a meta-analytic approach to increase precision and reduce Type I error, the three samples were analyzed individually and then averaged.

From Figure 7 (and Appendix D), it can be clearly seen that for the averaged data of Quizzes 1, 2, and 4, the mean and median test scores for the Only App Group (84.1%, 86.8%, respectively) were approximately 10% higher than those for the blended strategy group (74.6%, 78.2%) and about 30% higher than for the No App Group (56.8%, 56.7%). The amount of score dispersion (i.e., interquartile ranges) also differed considerably, with the Only App Group having much narrower ranges across the three quizzes, with an average spread of 15.7 (77.3–93), which is about a third narrower than both the No App Group (23.8; 45.3–69.1) and the blended Notes/App Group (24.6; 64.4–89). It is evident that the boxplots for the Only App Groups are consistently located higher than those for the other groups, though there are substantial overlaps with the Notes/App Group. The Quiz 2 data differs from the other two quizzes in that the Only App Group significantly outperformed not only the No App Group but also the Notes/App Group. The Quiz 4 data shows medians that are substantially lower than for the other quizzes, most likely due to the test difficulty because by Quiz 4, the learners were responsible for learning an accumulated total of over 500 words.

Figure 7

Boxplot Graphs for Each Quiz and Aggregated Data for All Three Quizzes



To test the data for statistically significant differences, the non-parametric Kruskal-Wallis chi-squared test was conducted (due to the non-normal distribution of values). As shown in Table 1, the Kruskal-Wallis chi-squared test for the three separate study strategy samples for Quizzes 1, 2, and 4 revealed statistically significant differences ($p < .05$) in the rankings of the scores for each of the quizzes, thereby allowing us to reject the null hypothesis and to accept the alternate hypothesis that the rankings for at least two out the three strategy groups are different; similarly, the Kruskal-Wallis test on the aggregated data (55.09, $df = 2$, $p < .000$) calculated a *p-value* much less than the alpha set at .05, thus allowing us to reject the null hypothesis with 95% confidence, and suggesting the acceptance of the alternative hypothesis that there is a statistically significant difference among at least two of the three study strategy groups. The Rank Sum Test that was next performed showed how much each strategy group differs from each other, and *r* coefficients were calculated as estimates of effect sizes.

Table 1

Results of the Non-parametric Kruskal-Wallis Chi-squared Test to Test for Significant Differences between the Three Study Strategies for Each Quiz

Data set	***H***	***Df***	***p-value***
Quiz 1	11.01	2	.004*
Quiz 2	32.73	2	.000***
Quiz 4	16.98	2	.000***
Avg: Quiz	20.20	2	p_{avg} = .001**
Agg: Quiz	55.09	2	.000***

*Note. *p < 0.05; **p < 0.01; ***p < 0.001*

Table 2

Results of the Wilcoxon Rank Sum Test (p-values with Benjamini-Hochberg Correction) to Specifically Compare Each Study Strategy for Each Quiz

Data Set	**Strategy Comparison**		
	C vs A	***C vs B***	***B vs A***
Quiz 1	.007**	.053	.035*
Quiz 2	.000***	.000***	.046*
Quiz 4	.000***	.085	.002**
Avg: Quiz	p_{avg} = .002**	p_{avg} =.046*	p_{avg} =.028*
Agg: Quiz	.000***	.000***	.000***

Note. A = No App; B = Notes and App; C = Only App
**p < 0.05; **p < 0.01; ***p < 0.001*

As could already be seen from the boxplot graphs of the data of the three groups across three quizzes, there are noticeable differences and similarities, which Table 2 quantifies by means of the Wilcoxon Rank Sum Test. The test identified statistical differences ($p < .05$) between seven out of nine study groups across the three quizzes; the two exceptions were Groups B (Notes/App) versus C (Only App) for quizzes 1 and 4 (p

= .053 and .085, respectively). In Table 3, the effect sizes for the statistically significant pairwise comparisons over the three quizzes indicate, on average, $medium_{PO}$ effect sizes of Group C over A (r = 0.41, 0.51, and 0.56, respectively; r_{avg} = 0.49) as compared to the $small_{PO}$ effect sizes of Group B over A for Quizzes 1 and 4 (r = 0.26 and 0.36; r_{avg} = 0.28); as mentioned above, Quiz 2 is noteworthy with Group C even having a $medium_{PO}$ effect size over Group B (r = 0.47). As for the aggregated data (i.e., Agg: Quiz) from Table 2 and Table 3, both Groups C and B significantly differed from Group A (Group C vs A: p_{agg} = 3.9e-10; Group B vs A: p_{agg} = .0000049) with respective $medium_{PO}$ and $small_{PO}$ effect sizes (r_{agg} = 0.50; and r_{agg} = 0.28), but even Group C significantly differed from Group B (p_{agg} = .00000066), with a $medium_{PO}$ effect size (r_{agg} = 0.28).

Table 3

Results of Wilcoxon Test to Compare the Three Vocabulary Studying Strategies (with Z Statistic and Effect Size r)

Data Set	**C vs A**		**C vs B**		**B vs A**	
	Z Score	***Effect Size (r)***	***Z score***	***Effect size (r)***	***Z score***	***Effect size (r)***
Quiz 1	-3.06	0.41^{++}	-1.94	0.18	-2.27	0.26^{+}
Quiz 2	-3.75	0.51^{++}	-5.10	0.47^{+}	-1.99	0.22
Quiz 4	-4.03	0.56^{++}	-1.72	7	-3.23	0.36^{+}
Avg: Quiz	z_{avg} = -3.60	$r_{avg} = 0.49^{++}$	z_{avg} = -2.92	$r_{avg} = 0.27^{+}$	z_{avg} = -2.50	$r_{avg} = 0.28^{+}$
Agg: Quiz	z_{agg} = -6.43	$r_{agg} = 0.50^{++}$	z_{agg} = -5.05	$r_{agg} = 0.28^{+}$	z_{agg} = -4.57	$r_{agg} = 0.30^{+}$

Note. A = No App; B = Notes and App; C = Only App

Effect Size (Plonsky & Oswald, 2014): $+small_{PO}$; $++moderate_{PO}$

Strategy Effectiveness and Efficiency

These analyses show that the Group C (Only App) is statistically different from the other groups. Specifically, there is 95% certainty that there is a statistically significant difference among the test-taking confidence rankings of the students who used the Only App strategy for studying vocabulary compared to the other two groups of students. In plain terms, the study strategy of only using the app generally led to higher test scores. Furthermore, the results suggest that using the app is efficient, if we consider studying efficiency as the relationship between test scores and hours of study (i.e., score/study hours). Although there are many uncertainties with the measurements taken in this study, the Only App Group seemed to use the most efficient strategy: as a group, they individually spent on average less time studying for the tests compared to the Notes/App Group (3.95 vs. 4.46 hrs), but achieved a 10% higher average score (84.1 vs 74.6%).

Fixed Effects Model Analysis

In order to determine whether other possible factors apart from study strategy may correlate with the dependent variable of test scores, post-quiz surveys were analyzed. A Fixed Effects (FE) Model was used to perform analyses on the following factors: study strategies (A, B, C), strategy switching (six variations), total time spent studying, total time using Quizlet, total number of separate times studying for the quiz, among others (Table 4). Using different factor combinations to judge the most efficacious combination of the smallest number of factors to account for the largest variance in scores (i.e., intercept value), the most parsimonious model included strategies, student code, quiz score, total time studying, and total time using Quizlet, with an intercept value of 86.32 and adjusted r^2 of 0.569 (residual standard error = 12.44 on 211 *Df* (Degrees of freedom); F = 4.114 on 155 and 211 *Df*; p = 2.2e-16). As Table 4 indicates, the studying strategy of Only App (p = .770) is no longer significantly associated with the variance in test scores and are replaced by "Quiz 4" and "total time using App". The negative value of the estimate for Quiz 4 (-4.05; p = .02) indicates the difficulty of this quiz compared to the previous ones, which should not be a surprise given that the learners were responsible for over 500 words by the time they had to take Quiz 4. As for the "total time using the App" (3.02; p = .005), this factor is statistically significant compared to the non-significant "total study time" (1.15; p = .24).

Table 4

Fixed Effects Model Measuring Moderating Effects of Factors on Test Scores

Factor	**Estimate**	**Std. Error**	***Df***	***t* Value**	**Significance *p* > *t***
	95% CI [2.5%, 97.5%]				
Intercept	86.32 [71.82, 100.81]	7.35	211	1.74	< .000***
S1.No App	-2.82426 [-10.82, 5.18]	4.06	211	-0.70	.49
S2.Only App	0.85 [-4.79, 6.49]	2.86	211	0.30	.77
Quiz 2 (vs Quiz 1 benchmark)	-2.48 [-5.99, 1.04]	1.78	211	-1.39	.17
Quiz 4 (vs Quiz 1 benchmark)	-4.05 [-7.46, -0.63]	1.73	211	-2.34	.02*
Total time studying	1.15 [-0.75, 3.04]	0.96	211	1.19	.24
Total time using App	3.02 [0.93, 5.11]	1.06	211	2.85	.005**

*Note. *p < 0.05; **p < 0.01; ***p < 0.001*

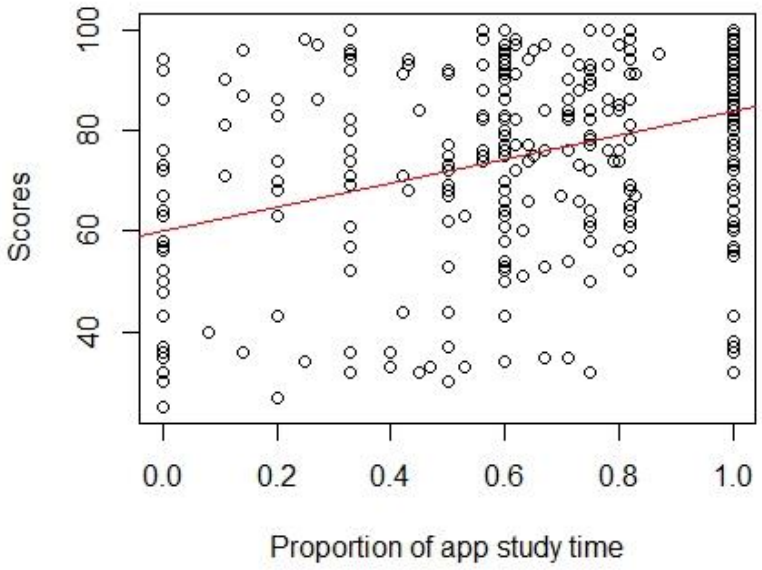
Follow Up FE Modeling

The results of another FE model with six strategy switching variations, study strategy, and quiz factors showed that none of the strategy switching factors were significant, thus providing evidence that they were not confounding factors (see Appendix E). On the other hand, this model showed that the Only App strategy had a significant positive effect (13.46, p = .022) on test scores, while Quiz 4 (again) had a significant negative effect (-5.37, p = .007). The model had an intercept value of 80.28 and adjusted r^2 of 0.51 (residual standard error = 13.23 on 207 *Df*; F = 3.42 on 159 and 207 *Df*; p = 2.2e-16). The results of the FE model further show that the first strategy (No App) was used as a comparison benchmark for the Notes/App and Only App strategies, with the positive coefficients of the two strategies indicating their relatively better test performances; the higher positive coefficient for the Only App strategy (13.46) was significant (p = .02) and indicates better test performances than for the Notes/App strategy (6.41). As for the effects of strategy switching, the only negative effect was switching from No App to Notes/App (A→B, -0.84) and the largest positive effect was, perhaps unsurprisingly, switching from No App to Only App (A→C, 19.11). Combining the relevant strategy coefficient with the strategy switching coefficient (e.g., C. Only App + A→C + B→C) can provide a summary of the strategy's total effect. In this sense, the total effect of the Only App strategy (13.46 + 19.11 + 5.04 = 37.61) was about 2.7 times greater than that of the Notes/App strategy (6.41 + 8.4 - 0.836 = 13.97).

When the "total study time using the App" (70.7%) was compared to the "total study time using Notes" (as a proportion of the total study time), it is clear that most students spent little time studying paper notes and wordlists. And as can be seen in Figure 8, the scatterplot shows a very wide dispersion, but most of the clustering occurs in the upper right quadrant, which represents higher test scores and app use. That is, the amount of app study time (as a proportion of the total study time) was positively correlated with test scores (r = 0.393). The results of both the FE model moderator analysis and correlation therefore reveals that the Only App strategy is not the main effect on the quiz scores but rather it is more generally the amount of time using the app.

Figure 8

Proportion of App Study Time (No App 0%-100% Only App) and Test Score Correlations



Discussion

Extent and Use of App

It is noteworthy how many learners used the app and the large amount of app usage, with more than 71% of the collective study-time across the three quizzes involving app-use. There are at least four reasons to explain this. Firstly, the students were highly motivated to improve their vocabulary knowledge. Secondly, Quizlet is a well-designed app with a simple interface and a variety of useful DF activities, which is why 212 out of 342 (62%) survey responses indicated that it was an effective tool for vocabulary learning (see Appendix C, question 9). The tool's effectiveness is also related to the learners' perceptions of the app being fun (24%) and creating opportunities to feel a sense of achievement and competition from completing activities and competing with classmates within and between classes for top scores on either the Match or Gravity game (25%); recall that the Match game was very popular and the second-most used activity over the four-month period (Figure 3). Thirdly, given the observation of the widespread "disconnection between technology and pedagogy" in many MALL studies (Burston, 2017, p. 2), effort was made to integrate the MALL component into the curriculum in four ways: (a) class time was devoted to introducing and demonstrating the app, as well as to getting the students to download the app and sign up for an account; (b) the tests were based on the app's vocabulary study sets, which were in turn based on the unit themes of the textbook; (c) use of the app was actively encouraged and recommended by the teacher; and (d) extra marks were awarded to students who completed app learning activities. The last two factors respectively accounted for 46% and 16% of survey responses about learner motivation to use the app. Finally, the students took advantage of the convenience of the app with more than half using the app outside of the home either anywhere, on the bus, or in school between classes (Figure 4), which reinforces Wu's (2015) observation that given the key role of repetition in language learning, MALL technologies need to be highly accessible and convenient.

Strategies and Factors Contributing to Higher Test Scores

While use of the app was encouraged for everyone, the students freely chose their own preferred studying method. Regarding the average amount of time studying over three quizzes, the No App Group appeared to be the least diligent, spending the least time on average at 2.54 hours per quiz (63.5 hrs / 25 people); the Only App Group was next at 3.95 hours (528.7 hrs / 134 people) and finally the Notes/App Group spent the most time on average at 4.39 hours (912.5 hrs / 208 people). Since the No App Group in this study was small in number and their dramatically lower test scores seem to generally represent less studious learners, they cannot be expected to represent students who typically use Notes to study, and for this reason, the No-App Group cannot credibly serve as a study strategy control nor benchmark. A more fruitful comparison, therefore, is between the other two larger-sized groups which both used the app.

As far as determining which study method was the most effective, the summary statistics of the three studying methods across three separate quizzes (as well as the averaged and aggregated data) indicated that the Only App strategy was more effective than either the No App or Notes/App strategy in terms of not only the means and medians for all three quizzes but also a tighter clustering of higher scores indicated by a narrower interquartile range of scores. The nonparametric Kruskal-Wallis Rank Sum Test further supported these findings. In all three quizzes, the Only App Group on average significantly outperformed the No App Group with a $medium_{PO}$ effect size (r = 0.49) and even significantly outperformed the Notes/App Group for Quiz 2 (p = .000001) with a $medium_{PO}$ effect size (r = 0.47). The Notes/App Group outperformed the No App Group on Quiz 1 and Quiz 4 with $small_{PO}$ effect sizes (r = 0.26, 0.36 respectively).

These effect sizes are comparable to meta-analyses that have showed the positive small effect of MALL on learning and for MALL in general (g = 0.52, or $r \cong 0.31$; Sung et al., 2016), and approaching a $medium_{PO}$ effect size MALL on vocabulary learning (g = 0.61 to 0.67, or $r \cong 0.37$–0.4; Mahdi, 2018). In this study, the groups that used the app (Groups B and C) clearly outperformed the No App Groups with effect sizes ranging from $small_{PO}$ (B vs A, r = 0.28) to $medium_{PO}$ (C vs A, r = 0.49), akin to those found by Mahdi (2018). However, as was mentioned above, these comparisons between the MALL Groups (B and C) and

non-MALL Group (A) are not entirely fair given the non-representative sample of Group A with its small sample size and low-scoring outliers that most likely indicated less motivated students who did not want to take the time to become familiar with the app. A more appropriate and novel comparison is between the two technologically-mediated strategies that used the app, though there is a dearth of such a strategy comparisons in the literature (See Elgort, 2018). Prior to the current study, the author assumed that the Notes/App strategy would be the most efficient studying method, specifically, studying the wordlists first and then using the app for review and consolidation. However, the Only App Group ended up with the most obvious and significant positive performance.

Moderators in the Fixed Effects Model

In order to better understand the differences between Groups B and C, an FE analysis was undertaken to determine what other moderating effects may be influencing the test scores of the strategy groups. Considering the noticeably higher average TOEIC scores in the Only App Group (median 750 vs. 665) and the greater proportion of males in the Only App Group (53% vs. 31%), both language ability and gender were considered to be possible moderators that inflated the Only App Group's scores. Other substantial raw number differences between Groups B and C suggested other factors, such as the average total number of hours of study per quiz (4.46 hrs vs. 3.95 hrs), the amount of time that Quizlet was used (2.63 hrs vs. 3.95 hrs), and even the devices used to run the app (e.g., smartphone usage: 89.6% vs. 58.9%).

Gender Differences?

There have been calls to conduct research into gender differences for mobile learning (Elaish et al., 2017), and gender did superficially seem to be a factor correlating with test scores and study strategy, with males using the app more than the females and preferring the Only App study strategy. This is consistent with meta-analytic studies on gender and use of educational ICT that have long noted the difference in attitude between males and females in terms of "belief" (i.e., the perceived usefulness of technology) and "self-efficacy" (i.e., confidence that one has sufficient abilities and skills to successfully employ information technologies) (Cai et al., 2017; Huang, 2013; Kay, 2008; Whitley, 1997). A recent example is Cai et al.'s (2017) meta-analysis of 50 studies from 1997 to 2014, which revealed that although attitude gaps between the genders are diminishing, there are still noticeable differences with males having more positive views toward beliefs and self-efficacy, with respectively *medium* ($g = 0.27$) and *small* effect sizes ($g = 0.18$), using Cohen's (1988) interpretive guidelines. However, the results of the FE model analysis showed that gender was not a moderator on the test scores. Similarly, the FE analysis failed to assign statistical significance to either TOEIC scores or devices used. Most importantly, running the FE model with different factors repeatedly revealed that neither the Notes/App nor Only App study strategies had a significant effect on the scores, especially when the factors total time studying and total time using app were added to the model.

"Use of App" as the Key Significant Moderator

As would be expected, the amount of study time influenced scores, but in a perhaps slightly surprising way. The FE analysis didn't find "total study time" to be a moderator on test scores ($p = .24$), but it did find "total time using the app" to be ($p = .005$). This is probably due to the very successful uptake of the app, with so many learners spending much of their study time using it. This finding perhaps explains why the non-parametric tests indicated that the Only App strategy was more successful, because most of the group members using this strategy utilized Quizlet more than those of the Notes/App Group (3.95 hrs vs 2.63 hrs); however, the FE analysis did not find the Only App study strategy to have a significant effect on scores, since many learners in the Notes/App Group also spent a lot of time using the app and received high scores. In the end, the finding that "total time using of app" was the most significant moderator shows that the strategy constructs of Only App and Notes/App were perhaps too simplistic and require a more fine-grained operationalization.

Limitations and Future Work

Given this study's non-experimental design, it has only established the positive correlation (not causal relationship) between the total time spent using the DF app to study vocabulary and scores received on vocabulary tests. Because the app was used by both the Notes/App and Only App Groups, it is difficult to tease out the relative effects of using the app compared to using notes or other specific strategies, like writing out words manually, verbal repetition, or forming an image of the meaning, any of which may have been used in conjunction with any of the three investigated study methods. In light of this, future studies will need a more sophisticated account of different strategies and learner choices of strategies in order to obtain a clearer picture of MALL for vocabulary learning and how it differs from a blended approach.

Furthermore, while strategy switching has been ruled out as a confounding factor, there are still several data issues in the present study that may challenge the validity of the findings, such as the different sized groups, small sample size of the No App Group, and the different contents of the quizzes. However, the findings of this study are useful in setting up hypotheses and designs for more controlled experiments that can employ repeated measures ANOVA tests and take these data issues into account. Future research should randomize groups and/or control the independent variable of study strategy to prevent apparently random switching and controlling other variables like student ability, which surprisingly appeared to be a non-significant moderator even though the Only App Group tended to have substantially higher TOEIC scores (see Figure 5). With better control and manipulation of variables, we will be more able to determine the causal influences of the app functions, total time studying, and total time using the app on receptive and/or productive vocabulary knowledge, to better quantify the effectiveness and efficiency of using the app to learn large amounts of vocabulary. Delayed posttests can also be performed to see if different methods or strategies have different memory effects.

Other limitations of the study include the generalizability of the findings due to the specific sample population of Taiwanese, post-graduate, and highly motivated students. Finally, a concern with all types of data collection and surveying using memory recall is the accuracy of the participants' memory. However, given the large sample size and repeated data collection, it is assumed that over- and under-estimations would cancel each other out across the 367 samples.

For app developers, if the users' active usage can be tracked for time, teachers and researchers can obtain useful information about how much effort was put into studying and can obtain data that is accurate for developing a "vocabulary learning efficiency metric" to compare different studying methods. At present, Quizlet can let teachers who have paid teacher accounts know which study activities were completed, however, it does not indicate the amount of time spent studying these activities and offers no information on how much time was spent on incomplete study activities. Given the short time and large amount of vocabulary that second language learners need to learn to achieve language proficiency, a fruitful line of future experimental research would be to strive for valid and reliable metrics of vocabulary studying efficiency by carefully controlling variables and more precisely tracking app usage time. The present study offers preliminary evidence to justify this kind of future research.

Conclusions

This study tracked 134 students over four months in the learning of more than 500 words in order to determine whether there were any discernible correlations between vocabulary learning strategies—including only notes and wordlists, combination of notes and a DF app, and only a DF app—and the scores from three vocabulary tests. The non-parametric tests showed that the Only App Group had significantly higher test scores than both the No App (Only Notes) Group and the blended Notes/App Group with $medium_{PO}$ and $small_{PO}$ effect sizes (r = 0.49 and 0.27, respectively). However, when a follow-up Fixed Effects model analysis was run to examine the moderating influences of gender, ability, time spent studying, time spent using the app, and frequency of studying, the Only App Group strategy became non-significant and was replaced by the factor "total time using the app" (p = .005) as the only statistically significant factor

associated with the test scores.

As the FE model revealed, learner characteristics and differences, such as gender, ability, and preferred device, had non-significant impacts on test performances, whereas only the amount of time using the app and (the most difficult) Quiz 4 had significant moderating effects on test scores. Contributing to the voluntary widespread adoption of the app was its mobile nature and convenience for anytime-anywhere vocabulary learning, as evidenced in 51% of the learners whose studying habits involved using use the app outside the home, anytime and anywhere (e.g., on the bus, or between classes in school). Another important reason explaining the successful uptake of the app was its integration into the curriculum, with the teacher's recommendation, it being the basis for frequent tests, having opportunities to earn extra marks by completing learning activities, and even being able to show off top scores to peers within and between classes.

The findings of this longitudinal study show the feasibility of using a well-designed mobile DF app like Quizlet for substantial outside-the-classroom vocabulary learning of more than 500 words over four months as an add-on component to a regular (business) English course. In fact, it is more than just feasible, as evidenced by this study, and the evidence even suggests that use of the app alone was the most effective way to study large amounts of vocabulary for vocabulary tests in terms of requiring less study time to get higher scores. Given the autonomy and freedom that the learners were given to choose their studying method, their overwhelming use of the app seems to validate it for both genders and all proficiency levels as a "one-stop learning shop" that can efficiently facilitate receptive vocabulary acquisition. The clear correlation between the amount of app study time and higher test scores is testimony to the app's effectiveness in vocabulary learning, especially when compared to the mechanical grind of rote memorization with wordlists and rote copying. Apps like Quizlet can offer a variety of useful activities and entertaining games that appear to be more time-saving and efficient for memorizing vocabulary than other strategies, and can produce motivational effects by providing users with fun, achievement and competition, which can be a rich add-on component to any language learning course.

Acknowledgements

I would like to thank the reviewers who offered very useful feedback and additional literature recommendations. The paper is greatly improved from their comments and suggestions. I am also indebted to Dr. Yicheng Kao for introducing me to and helping me with the Fixed Effects model.

References

Afzali, P., Shabani, S., Basir, Z., & Ramazani, M. (2017). Mobile-assisted vocabulary learning: A review study. *Advances in Language and Literary Studies*, *8*(2), 190–195.

Agca, R. K., & Özdemir, S. (2013). Foreign language vocabulary learning with mobile technologies. *Procedia-Social and Behavioral Sciences*, *83*, 781–785. https://doi.org/10.1016/j.sbspro.2013.06.147

Alahmadia, A., Shanka, C., & Foltzb, A. (2018). Vocabulary learning strategies and vocabulary size: Insights from educational level and learner styles. *Vocabulary Learning and Instruction*, *7*(1), 1–21. https://doi.org/10.7820/vli.v07.1.alahmadi

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. SAS Institute.

Amer, M. (2010). *Idiomobile for learners of English: A study of learners' usage of a mobile learning application for learning idioms and collocations* (Publication No. 3413155)[Doctoral dissertation, Indiana University of Pennsylvania]. ProQuest Dissertations Publishing.

Anaraki, F. (2009). A flash-based mobile learning system for learning English as a second language. In *Proceedings of the International Conference on Computer Engineering and Technology ICCET, Vol. 1* (pp. 400-404). IEEE. https://doi.org/10.1109/ICCET.2009.183

Azabdaftari, B., & Mozaheb, A. M. (2012). Comparing vocabulary learning of EFL learning by using two different strategies: Mobile learning vs. flashcards. *The EUROCALL Review*, *20*(2), 48–59. https://doi.org/10.4995/eurocall.2012.11377

Başoğlu, E. B., & Akdemir, O. (2010). A comparison of undergraduate students' English vocabulary learning: Using mobile phones and flash cards. *Turkish Online Journal of Educational Technology-TOJET*, *9*(3), 1–7.

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*, *27*(1), 4–20. https://doi.org/10.1017/S0958344014000159

Burston, J. (2017). MALL: Global prospects and local implementation. *CALL-EJ*, *18*(1), 1–8. http://callej.org/journal/18-1.html

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers & Education*, *105*, 1–13. https://doi.org/10.1016/j.compedu.2016.11.003

Çelik, Ö., & Yavuz, F. (2018). An extensive review of literature on teaching vocabulary through mobile applications. *Bilecik Şeyh Edebali Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *3*, 56–91. https://doi.org/10.33905/bseusbed.393947

Chien, C. W. (2015). Analysis the effectiveness of three online vocabulary flashcard websites on L2 learners' level of lexical knowledge. *English Language Teaching*, *8*(5), 111–121. https://doi.org/10.5539/elt.v8n5p111

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cook, L., & Mayer, R. (1983). Reading strategies training for meaningful learning from prose. In M. Pressley & J. Levin (Eds.), *Cognitive Strategy Research: Educational Applications* (pp. 87–131). Springer. https://doi.org/10.1007/978-1-4612-5519-2_4

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge. https://doi.org/10.4324/9781315708607

Davie, N., & Hilber, T. (2015, March 14–16). *Mobile-Assisted language learning: Student attitudes to using smartphones to learn English vocabulary* [Paper presentation]. 11th International Association for Development of the Information Society (IADIS) International Conference on Mobile Learning Madeira, Portugal.

Dizon, G. (2016). Quizlet in the EFL classroom: Enhancing academic vocabulary acquisition of Japanese university students. *Teaching English with Technology*, *16*(2), 40–56.

Dizon, G., & Tang, D. (2017). Comparing the efficacy of digital flashcards versus paper flashcards to improve receptive and productive L2 vocabulary. *The EuroCALL Review*, *25*(1), 3–15. https://doi.org/10.4995/eurocall.2017.6964

Duffy, J. (2020, Dec. 22). The best free language-learning apps of 2021. *PC Magazine*. https://uk.pcmag.com/migrated-45785-education/116058/the-best-free-language-learning-apps

Elaish, M. M., Shuib, L., Ghani, N. A., Yadegaridehkordi, E., & Alaa, M. (2017). Mobile learning for English language acquisition: Taxonomy, challenges, and recommendations. *IEEE Access*, *5*, 19033–19047. https://doi.org/10.1109/ACCESS.2017.2749541

Elgort, I. (2018). Technology-mediated second language vocabulary development: A review of trends in research methodology. *CALICO Journal*, *35*(1), 1–29.

Fisher, T., Pemberton, R., Sharples, M., Ogata, H., Uosaki, N., Edmonds, P., Hull, A., & Tschorn, P. (2009). Mobile learning of vocabulary from reading novels: A comparison of three modes. In D. Metcalf, A. Hamilton, & C. Graffeo (Eds.), *Proceedings of 8th World Conference on Mobile and Contextual Learning* (pp. 191–194). University of Central Florida.

Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language Learning Journal*, *36*(2), 239–248. https://doi.org/10.1080/09571730802390759

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education*. McGraw-Hill.

Frohberg, D., Göth, C., & Schwabe, G. (2009). Mobile learning projects: A critical analysis of the state of the art. *Journal of Computer Assisted Learning*, *25*(4), 307–331. https://doi.org/10.1111/j.1365-2729.2009.00315.x

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education*, *28*(1), 1–35. https://doi.org/10.1007/s10212-011-0097-y

Jackson III, D. B. (2015). A targeted role for L1 in L2 vocabulary acquisition with mobile learning technology. *Perspectives (TESOL Arabia)*, *23*(1), 6–11.

Jiménez Catalán, R. M. (2003). Sex differences in L2 vocabulary learning strategies. *International Journal of Applied Linguistics*, *13*(1), 54–77. https://doi.org/10.1111/1473-4192.00037

Kay, R. H. (2008). Exploring gender differences in computer-related behavior: Past, present, and future. In T. T. Chen & I. Chen (Eds.), *Social information technology: Connecting society and cultural issues* (pp. 12–30). Information Science Reference. https://doi.org/10.4018/978-1-59904-774-4.ch002

Kiliçkaya, F., & Krajka, J. (2010). Comparative usefulness of online and traditional vocabulary learning. *TOJET: The Turkish Online Journal of Educational Technology*, *9*(2), 55–63.

Klímová, B. (2018). Mobile phones and/or smartphones and their apps for teaching English as a foreign language. *Education and Information Technologies*, *23*(3), 1091–1099. https://doi.org/10.1007/s10639-017-9655-5

Lin, J. J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *32*(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

Liu, G. Z., Lu, H. C., & Lai, C. T. (2014). Towards the construction of a field: The developments and implications of mobile assisted language learning (MALL). *Digital Scholarship in the Humanities*, *31*(1), 164–180. https://doi.org/10.1093/llc/fqu070

Mahdi, H. S. (2018). Effectiveness of mobile devices on vocabulary learning: A meta-analysis. *Journal of Educational Computing Research*, *56*(1), 134–154. https://doi.org/10.1177/0735633117698826

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Heinle & Heinle Publishers.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Heinle.

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Descriptive, acquisition and pedagogy* (pp. 6–19). Cambridge University Press.

Nation, I. S. P., & Yamamoto, A. (2012). Applying the four strands to language learning. *International Journal of Innovation in English Language Teaching and Research*, *1*(2), 167–182. https://doi.org/10.26686/wgtn.12552020

NewZoo. (2018, May). *Insights: Top 50 countries/markets by smartphone users and penetration*. https://newzoo.com/insights/rankings/top-countries-by-smartphone-penetration-and-users/

Nikoopour, J., & Kazemi, A. (2014). Vocabulary learning through digitized & non-digitized flashcards delivery. *Procedia–Social and Behavioral Sciences*, *98*, 1366–1373. https://doi.org/10.1016/j.sbspro.2014.03.554

Ou-Yang, F. C., & Wu, W. C. V. (2017). Using mixed-modality vocabulary learning on mobile devices: Design and evaluation. *Journal of Educational Computing Research*, *54*(8), 1043–1069. https://doi.org/10.1177/0735633116648170

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. Newbury House.

Pavičić Takač, V. (2008). *Vocabulary learning strategies and foreign language acquisition*. Multilingual Matters.

Plonsky, L., & Oswald, F. L. (2014). How big is "big?": Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Quizlet. (2022, February 24). *About Quizlet*. https://quizlet.com/mission

R Core Team. (2020). *R: A language and environment for statistical computing* (version 3.5.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). Russell Sage Foundation.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Descriptive, acquisition and pedagogy* (pp. 199–227). Cambridge University Press.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/s0261444812000018

Song, Y., & Fox, R. (2008). Using PDA for undergraduate student incidental vocabulary testing. *ReCALL*, *20*(3), 290–314. https://doi.org/10.1017/s0958344008000438

Stockwell, G. (2007). Vocabulary on the move: Investigating an intelligent mobile phone-based vocabulary tutor. *Computer Assisted Language Learning*, *20*(4), 365–383. https://doi.org/10.1080/09588220701745817

Stockwell, G. (2010). Using mobile phones for vocabulary activities: Examining the effect of the platform. *Language Learning & Technology*, *14*(2), 95–110. https://doi.org/10125/44216

Sung, Y. T., Chang, K. E., & Liu, T. C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275. https://doi.org/10.1016/j.compedu.2015.11.008

Sung, Y. T., Chang, K. E., & Yang, J. M. (2015). How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review*, *16*, 68–84. https://doi.org/10.1016/j.edurev.2015.09.001

Taj, I. H., Sulan, N., Sipra, M., & Ahmad, W. (2016). Impact of mobile assisted language learning (MALL) on EFL: A meta-analysis. *Advances in Language and Literary Studies*, *7*(2), 76–83. https://doi.org/10.7575/aiac.alls.v.7n.2p.76

Tosun, S. (2015). The effects of blended learning on EFL students' vocabulary enhancement. *Procedia–Social and Behavioral Sciences*, *199*, 641–647. https://doi.org/10.1016/j.sbspro.2015.07.592

Turner, J. L. (2014). *Using statistics in small-scale language education research: Focus on non-parametric data*. Routledge. https://doi.org/10.4324/9780203526927

Whitley Jr., B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior*, *13*(1), 1–22. https://doi.org/10.1016/s0747-5632(96)00026-x

Wu, Q. (2015). Pulling mobile assisted language learning (MALL) into the mainstream: MALL in broad practice. *PloS One*, *10*(5), e0128762. https://doi.org/10.1371/journal.pone.0128762

APPENDIX A. Effect Size Guidelines Using Plonsky and Oswald (2014) as a Benchmark

Categories	Plonsky and Oswald (2014)	Cohen (1988)	Plonsky and Oswald (2014)	Cohen (1988)
	r	***r***	***d***	***d***
small	.25	.1	.4	.2
medium	.4	.3	.7	.5
large	.6	.5	1.0	.8

APPENDIX B. Sample Items on a Vocabulary Quiz

Sample items testing receptive vocabulary knowledge:

	a. involving a lot of complicated rules, details, and processes
fickle ___	b. producing powerful feelings or strong, clear images in the mind
prominent ___	c. serving as a temporary or expedient means, especially during an emergency
	d. very well-known and important
vivid ___	e. deprived of one's job because it is no longer necessary for efficient operation
bureaucratic ___	f. unfriendly and not agreeing with something
	g. always changing in purpose, affections or mood

Sample items testing productive vocabulary knowledge:

_____ amount paid ahead of time to secure a reservation

_____ to commence officially or formally initiate

_____ the process of getting supplies

_____ to carry, convey, or drag

APPENDIX C. Post-test Survey Questions and Survey Results

Question	Option 1	Option 2	Option 3	Option 4	Option 5	Option6	Option 7	Option 8
1. I think vocabulary is an important part of language learning.	Strongly disagree 0.5%	Disagree 0%	Neither agree / disagree 1.6%	Agree 19%	Strongly agree 79%	-	-	-
2. I have a strong desire to improve my vocabulary ability.	Strongly disagree 1%	Disagree 1%	Neither agree / disagree 3.7%	Agree 30.5%	Strongly agree 63.7%	-	-	-
3. How much time did you spend studying for the quiz?	0-1 hr 2.4%	1-2 hr 11.4%	2-3 hr 23.9%	3-4 hr 10.4%	4-5 hr 19.6%	5-6 hr 9.8%	6+ hr 22.3%	-
4. How much time did you use Quizlet to study for the quiz?	I didn't use it 6.8%	0-1 hr 8.2%	1-2 hr 23.9%	2-3 hr 19.6%	3-4 hr 25%	4-5 hr 15.5%	5-6 hr 6.8%	6+ hr 7.6%
5. How many separate TIMES did you study before the test?	1-2 24.5%	3-4 36.8%	5-6 18.8%	7-8 10.4%	9-10 6%	10+ 3%	-	-
6. What was your vocabulary studying strategy?	No App, only notes/ wordlists 7%	Notes/word-lists and App 51%	Only App 42%	-	-	-	-	-
7. What Quizlet activities did you use? (N=342)	Flashcards 34.6%	Learn 64%	Speller 22.6%	Write 21.2%	Test 26.7%	Match game 35%	Gravity game 15%	-
8. On what device did you primarily use Quizlet?	I didn't use Quizlet 7%	Smart-phone 62%	Tablet pc 6%	Notebook pc 21%	Desktop pc 3.5%	-	-	-
9. If you used Quizlet, what was your **main** motivation? (check 1-2; N=342)	Challenge and achieve-ment of top game score 25.6%	For extra marks 15.8%	Fun activities, like the scatter game 24.6%	Teacher's encour-agement & recomm-endation 46%	Quizlet is an effective tool for learning 62%	-	-	-
10. Where did you usually use Quizlet? (N=342)	Home 42.9%	Between classes at school 2.6%	On the bus, MRT, etc. 26%	Anywhere, anytime I had time 28%	-	-	-	-

Note. N = 367 for all of the questions, except for 7, 9, and 10 (n = 342)

APPENDIX D. Tabulated Summary of Data for the Three Groups Across Three Quizzes and Aggregated (gender, *N*, means, medians, ranges)

Quiz *N (% male)*	**Group A: No App** *N (% male)* *Mean* *Median + (Interquart. range)*	**Group B: Notes/App** *N (% male)* *Mean* *Median + (Interquart. range)*	**Group C: Only App** *N (% male)* *Mean* *Median + (Interquart. range)*
Quiz 1	8 (12.5%)	67 (32.8%)	49 (51%)
N = 124 (39%)	58.8	76.5	83.5
	62.0 (50.8-68.3)	82 (64.5-93)	84 (74-95)
Quiz 2	9 (44%)	70 (28.6%)	46 (52%)
N = 125 (38%)	58.2	74.2	88.9
	52 (48-76)	76.5 (66.2-86)	94.5 (84-96)
Quiz 4	13 (30.1%)	66 (31.8%)	39 (56%)
N = 118 (40%)	53.3	73.1	79.8
	56 (37-63)	76 (62.5-88)	82 (74-88)
Quizzes 1, 2, 4 Averages	10 (28.9%)	67.7 (31.1%)	44.7 (53%)
N = 122.3 (39%)	56.8	74.6	84.1
	56.7 (45.3-69.1)	78.2 (64.4-89)	86.8 (77.3-93)
Aggregated data			
N = 367 (39%)	56.27	74.6	84.3
	56.5 (38.5-66.75)	77 (64-91)	87 (77-95)

APPENDIX E. Fixed Effects Model Showing Non-significant Effects for Strategy Switching

Factor	Estimate	Std. Error	*Df*	*t* Value	Significance *p* > t
	95% CI [2.5%, 97.5%]				
Intercept	80.28 [61.44, 99.11]	9.55	207	8.40	< .000***
B. Notes/App	6.41 [-4.53, 17.35]	5.55	207	1.16	.25
C. Only App	13.46 [1.97, 24.96]	5.83	207	2.31	.022*
Quiz 2 (vs Quiz 1 benchmark)	-0.80 [-4.63, 3.02]	1.94	207	-0.41	.68
Quiz 4 (vs Quiz 1 benchmark)	-5.71 [-9.61, -1.81]	1.98	207	-2.89	.004**
B→A	1.02 [-12.30, 14.34]	6.76	207	0.15	.88
C→A	5.95 [-7.16, 19.07]	6.65	207	0.90	.37
A→B	-0.84 [-14.18, 12.51]	6.77	207	-0.12	.90
C→B	8.40 [-0.07, 6.879]	4.30	207	1.95	.052
A→C	19.11 [-1.74, 39.96]	10.58	207	1.81	.07
B→C	5.04 [-3.45, 13.53]	4.31	207	1.17	.24

*Note. *p < .05; **p < .01; ***p < .001.*

About the Author

Nigel P. Daly has been teaching and researching English as a Foreign Language in Taiwan for over 23 years. His research interests are in Rasch modeling and English as a Lingua Franca as it specifically applies to language testing, research and business writing, and social media use.

E-mail: ndaly@hotmail.com