A Determinant of Optimal and Inhibited Mobile Language Learning Activity

Quiz Level Length

https://doi.org/10.3991/ijim.v17i10.38417

Jason Byrne
Toyo University, Tokyo, Japan
Tokyo Denki University, Hatoyama, Japan
byrne@toyo.jp

Abstract—Learner analytic exploratory research, concerned with mobile app learner survivability, was conducted. An app was developed to indicate productivity of quiz app usage for global mobile assisted language learning (MALL). The research focused on inhibiting and optimising effects of level length on total questions answered. Specifically, it aimed to answer the question: how many questions per level leads to the highest total unique questions being answered? The research was conducted within nine-day cohort timeframes and included three phases: a small-scale pilot study to establish parameters, an exploratory stage undertaken within the parameters and finally the use of a quadratic regression predictive model. The data was collected using Google Analytics and Google Firebase. The null hypothesis (H0) was rejected. A one-way analysis of variance found a statistically significant difference in the mean productivity of the lengths. Results of a Tukey post hoc test ($p < .05$) suggests question sets with less than eight questions, or more than 15 questions, appear to inhibit MALL autonomous learning. Optimal level question sets appear to be between lengths 8-14. The results visually encapsulated by a quadratic regression model broadly support H1 and H2. Set 12 is the statistically most significant optimal load. Gains of 107% are reported for switching from suboptimal to optimal approaches. The major conclusion of the research is that quiz length strongly affects quantity of work completed. A preliminary time-based model of optimal length has been provided, to broaden measurement opportunity, as the findings could potentially cross-apply to professional environments and other activity types.

Keywords—activity theory, behaviourism, EFL, facilitation of m-learning, games, learner analytics, mobile language learning, productive learning, technology enhanced learning

1 Introduction

This exploratory behavioural paper investigates one potential lever of technological influence on the ability of students to control their own learning. The study looks

at the effect of quiz level length on learner output in a mobile language learning (MALL) app. The paper hypothesises that m-learning quiz level length is neither benign nor neutral. This is important research, as potentially a simple tweak to level length, could significantly improve, or impede, the quantity of student work completed. In addition, quiz level length effect, could be distorting the reporting of other more high-profile behavioural design choices, such as gamification and digital nudging. The research asks and answers the question, is there an optimal quiz level length that leads to more work being completed?

Recently, there has been greater emphasis on educational games and gamification in the classroom [1]–[5]. This has partially been shaped by the experiences of the COVID-19 pandemic [6]–[9], and increased ICT integration into educational practice [10]. Additionally, home usage has been increasingly facilitated by parents [11]. However, mobile assisted language learning (MALL) seems to be predominantly autonomous self-study activity [12]. According to a recent study [13], diverse informal vocabulary learning is an important and increasing trend. This likely implies most MALL users are driven by their own goals and their own intention. This assertion is supported by a study [14] showing MALL usage is truly 24/7. Users keep their own hours. This strength of individual intention further implies the potential role of gamification to leverage it. There has been a wealth of recent research into the merits of gamification to stimulate learning (see e.g., [15]–[19]). On the other hand, personal intention also implies that any inhibitor to MALL based self-study could have a serious negative impact on learning effects. That said, gamification can be used to improve educational outcomes [20]. One of the primary tenants of gamification is the concept of user autonomy. The user volunteering to participate is central to why gamification works [21], [22]. This study looked at users who volunteered to install an English as a foreign language (EFL) quiz app and then chose to level up and play with a sense of intention. But at a very fundamental level, can question load inhibit this intention? In other words, if the wrong balance is set, does it inhibit the survivability of learning and gamification? And conversely, by finding the optimum number of questions to deliver as a set, is it possible to maximise learning undertaken?

1.1 Theoretical framework

Flexible and iterative design-based research (DBR) performed within a real mobile educational ecosystem [23], [24]. The research was framed by activity theory and analytically approached from the perspective of behavioural psychology. Activity theory [25], [26] is fundamentally based on the concept of actors using a mediating tool to reach an objective [27], [28]. In this case, the higher-level objective is to learn English, the mediating artefact is a MALL app, and the actors are users of the app. The research was focused on the use of the tool to mediate individual learning as proposed by Vygotsky [27] and furthered by Leont′ev [28]. In Leont′ev's second generation activity model, activity is built on a chain of actions [28], with each action comprising a subset of operations. The operations are the how of action completion [29], and at the lowest level can be automatic responses to conditions [25]. In the third generation, Engeström's activity system [30] added a layer of social and communal

complexity, suitable for the classroom, but that was largely outside the boundaries of this research project. In fact, this research is focused on an object-orientated lower level of abstraction, in the sense that the tool is also an object, the outcome of tool creation activity. Therefore, the tool is also the product of action and operation. This research focuses on one operation that is a modular component of the tool that in turn at a higher level of abstraction mediated learner outcomes. In terms of operation, the research was interested in a nuanced understanding of how a tool can be adjusted to produce magnified enablement and/or constraint. Specifically, how the quantity of actor usage of the mediating artefact is causally determined by quiz level length. The interplay of the operation's underlying interaction with conditions and ultimate effect on the user activity, appeared to lend itself to analysis by behaviourism [31]–[35]. The analysis was based on a subconscious interaction of stimulus and response [31], [32] and Thorndike's three laws of learning [36].

1.2 Hypotheses

The following hypotheses are posited for an autonomous learner orientated EFL quiz app.

- H_0 The number of questions per level has no impact on progressive work completed.
- H_1 The optimal payload of questions per level leads to optimal progressive new work being completed.
- H_2 A sub-optimal quiz load will inhibit autonomous learning.

In order to test the hypotheses, an app was created. The app was precisely the same for all users with one key difference. The tested difference was the number of questions per game level. A review of research using EFL quizzes found the length of quizzes to generally range from 2-15 questions (e.g., [37]–[39]). Applied computer science research has been previously undertaken on optimal question set size in terms of user survivability on crowdsourcing platforms [40]. Employing a user longevity task allocation strategy, data collection was increased by up to 117.8%. The study found the end of a given question set often signifies an exit point for a user session. In other words, users will generally make it to the end of a quiz, if they know how long the quiz is and it is not unreasonable, but then will stop [40]. The other take away is that the highest mean average questions answered were for the longest question sets. A question set of 50 led to an average of 25 answers. In other words, stretching people can improve gains. But users typically do not complete the set of 50. A different study [41] found EFL quiz games were generally played for between 2 mins 45 seconds and 4 minutes 15 seconds. It seems unlikely that more than 15 questions would be answered in under 5 minutes. Therefore, this suggested that the best performing set might be under 15 questions, as EFL researchers are tending to use. In contrast, the earlier mentioned findings [40] suggest 25 questions might be a possible optimal length. Although a 25-question user average did seem to be based on over-stretching users to the point of failure, which may not be conducive to a learning environment. However, to ensure bias was not being front-loaded into the research design, users

were provided with between one and thirty questions per level. It was decided that if the higher question groupings performed well, then the research would be extended to even higher numbers. If they performed poorly, then the emphasis would be on narrowing towards an optimum.

2 Methods

2.1 Research instrumentation

A simple quiz app was developed using Unity software [42] and a quiz game template kit [43]. Firebase Analytics [44] and Google Analytics [45] were used to collect data from the app. This is achieved by inserting small code snippets into the app activity class code to monitor user actions. The pilot study made use of the analytics platforms' real-time data on the web-based dashboard. The exploratory data was collected and downloaded as a csv file. The app itself involves the answering of English as a foreign language (EFL), multiple-choice quiz questions. Each question provides three choices, the correct answer and two distractors. The questions are set in a presequenced order. If the user makes it through a stage and does not run out of lives, then they will never see the same question again. They will progress or move forwards into new question areas. If they run out of lives, they start from the beginning and follow the same sequence. Initially, users are provided with three lives to familiarise themselves with the game. However, from the second game they are provided with 100 lives, meaning there was no reason to repeat a stage, unless they opted to stop and exited the game mid-stage.

2.2 Sampling procedure

A non-probability sampling procedure was implemented, the participants forming a voluntary response opt-in group. This approach was practical and allowed for broad exploration. A random sample would have required for the same research design, the recruitment of a cohort of 900 users and the ability to provide them 900 devices. Realistically, the research design would have become more limited in scope, but would have still required one device per user, possibly 100 devices. The researcher did not have access to such a large number of devices. If the research had used users' personal devices, then the sample would have once again become a voluntary sample, and this would have raised additional issues, such as a potential power imbalance between likely student "volunteer" users and the teacher-researcher. Consequently, the app was published as an open test on Google Play. This means that the app was not formally published but was openly accessible. The users were anonymously recruited through Google Ads. This is a novel participant recruitment approach that has been used in recent times [46], [47]. The participants were anonymous to the researcher. No personally identifiable data was collected. The sample was derived from a population of Android MALL users. The parameters of the population were largely defined by researcher ability to pay and the Google Ads AI system, in this respect it was a

convenience sample. For clarification, payment was to Google Ads for showing the ads, they were shown on Google search pages, YouTube, and other third-party properties. The actual users received no payment.

2.3 Participants

Pilot data was not segmented due to small cohort sizes. Its function was primarily to check that the randomisation code was working and to get an early indication of how the lengths were inhibiting (or not) the user activity. During the primary research period, the exploratory stage, 2174 users who reached the second level of a quiz were identified by both age and gender. Their device settings enabled this information to be shared. This was the app sample that had met all pre-conditions for selection (known age, gender, and reached level two). 56% of the sample were women and 41% of the sample were between 18-24 (see Table 1).

Table 1. Age & gender

	Raw			**Age by Gender**			**Gender by Age**		
	Female	***Male***		***Female***	***Male***		***Female***	***Male***	***Total***
18-24	574	326	18-24	0.64	0.36	18-24	0.47	0.34	0.41
25-34	227	193	25-34	0.54	0.46	25-34	0.19	0.2	0.19
35-44	137	138	35-44	0.5	0.5	35-44	0.11	0.14	0.13
45-54	107	120	45-54	0.47	0.53	45-54	0.09	0.13	0.1
55-54	89	101	55-54	0.47	0.53	55-54	0.07	0.11	0.09
65+	84	78	65+	0.52	0.48	65+	0.07	0.08	0.07
N=2174	1218	956	All Ages	0.56	0.44		1.0	1.0	

The users reaching a second level was deemed important, as it screened out users who had downloaded the app and then never played. It improved the sense of learning intention and focused the research sample on genuine learners. The app was used internationally, users made it to at least the second level in 151 countries (or territories). Approximately 75.2% of usage in the largest age category 18-24, that met the demographic criteria (age & gender), was located in 21 countries. The other 24.8% of usage it is implied came from the other 130 countries, but which countries are unclear. However, it is known that the smallest recorded entry in the data is 47 events. Therefore, we can assume the individual country event data entries for the final 24.8% of the 18-24 age category are probably less than 47. This implies at least a further 42 countries represented in the 18-24 category which represent 0.45 of all usage. In the other age categories, due to anonymity thresholds, far less countries are disclosed, and explanation would be less meaningful, but all countries that are listed are among the 21 in Table 2 for 18-24-year-olds. The key point is that the users are dispersed globally with a particularly strong presence in Africa, the Middle East, and Asia. Myanmar has the strongest usage pattern and accounts for 10.6% of all known usage. Egypt accounts for 7.2% and Algeria 5.3%. Iraq (4.3%), Bangladesh (3.4%) and Pakistan (2.1%) are the only others above 2%.

Table 2. Top country usage for 18-24-year-olds

Myanmar (Burma)	1600	Ethiopia	217	Colombia	158	Uzbekistan	100
Egypt	682	Morocco	207	Jordan	145	Azerbaijan	92
Bangladesh	437	Tunisia	207	India	140	Turkey	83
Iraq	411	Somalia	195	Lebanon	131	Sudan	69
Pakistan	374	Palestine	177	Vietnam	125	Nepal	53
Algeria	276						

The population was initially randomly assigned question sets 1-30 during the pilot stage. But the research narrowed and only 14 sets were recorded in detail. This largely explains why 59.3% of the population sample is used. Only a snapshot of the activity of the potential qualifying app population (N=2174) are included in the experimental windows. The main cohort participants (n=1291) found in the exploratory stage, entered, and completed at least one level in the recorded cohort experiments, at some point during the nine-day windows. The majority entered in the first 48 hours of the window, as this was the recruiting period when the ads were live, but there was no way to prevent others joining or controlling when a recruited participant would complete the requirements of recruitment. This was not a lab experiment; the social environment could not be controlled. Furthermore, the research is focused on the general impact of question length on learning and not segmented by age, gender, or country. However, for reference, the cohort is slightly skewed towards women (56%) and heavily skewed towards younger adults (41% < 25 years of age, 60% < 35 years of age). In terms of actual levels played and passed, women account for 64% of activity (see Table 3), suggesting they were more active than men. These factors need to be considered when extrapolating from the data.

Table 3. Usage by age & gender

	Raw			**Age by Gender**			**Gender by Age**		
	Female	***Male***		***Female***	***Male***		***Female***	***Male***	***Total***
18-24	5237	2584	18-24	0.67	0.33	18-24	0.47	0.41	0.45
25-34	1622	1253	25-34	0.56	0.44	25-34	0.15	0.2	0.16
35-44	1110	927	35-44	0.54	0.46	35-44	0.1	0.15	0.12
45-54	1018	531	45-54	0.66	0.34	45-54	0.09	0.08	0.09
55-54	1343	627	55-54	0.68	0.32	55-54	0.12	0.1	0.11
65+	783	408	65+	0.66	0.34	65+	0.07	0.06	0.07
Levels	11113	6330	Total	0.64	0.36				

2.4 Participant consent

The users were invited and opted into the research. This involved two steps. Firstly, they would have clicked on the advert for an English quiz game. Approximately 2.5% of ad viewers clicked. Secondly, once at the quiz game store page, they would have chosen to install the app and 47% did elect to install the app. In other words,

roughly 85 potential participants were invited for every participant who elected to install the app. The app store page clearly stated that the app was collecting data for academic research purposes. This was also stated in the privacy page and a link was included inside the app. The data itself was collected using Google Firebase (GF) and Google Analytics (GA). It was anonymised at source. Google Analytics (GA) includes data thresholds. Essentially, GA do not release the data to collectors until the app users` anonymity is assured. Furthermore, GA automatically delete any identifying information, such as age and gender segmentation after 90 days, to comply with international standards, for example, EU general data protection regulation (GDPR). For clarification, no attempt was made to uncover user identities and the data was deleted from GF and GA once the research period was concluded. Also importantly, the study is in accordance with the Helsinki Declaration, and comparable ethical standards. In terms of human research as defined in the Helsinki Declaration [48], [49] and Belmont Report [50], no individualised data was collected. Since no individual data was collected, the risk of unintentional harm to the participant was greatly reduced. In addition, no direct physical contact was ever made with the participants. The project was essentially a form of observational research in a digital social context. Furthermore, the research data is based on the playing of an educational app targeted at, and exclusively collecting data on, adult users. US federal policy 45 CFR 46.104-d-3-ii [51] cites adult online game activity as an example of benign research activity in the context of education. Given the online nature of the research and the fact American analytic data platforms were being used to collect participant data, US regulatory standards seemed particularly relevant as a source of ethical guidance.

2.5 Random group placement

The installed users were initially placed into 30 groups. The placement was effectively random. Inserted in the app code was an initial randomisation set up function. The number of quiz questions per level for each user was set for the duration of the project stage at between one and thirty questions. It can be argued that randomly produced numbers are essentially pseudo-random rather than truly random. However, randomness is effectively achieved by inputting the time stamp of the function call. The time the user chose to open the app could not be controlled by the researcher, and the comparative selection of time across the global cohort of users was unknowable to the user, consequently this assured unpredictability (randomness) in group selection. It quickly became apparent that randomly collecting data from 30 groups would be expensive, this supports the previous feasibility study findings [46]. For example, creating 30 groups with a minimum of 30 participants (minimum n=900), based on age, gender, country and who had returned for a second session, was estimated to require the recruitment of over 15,000 users. The estimate is based on achieving minimum cohorts of 30 participants in each of 30 groups (est. n=900). But working backwards, only 33% of users will return for a second session (est. n=2700). Some groups will return at much lower rates (approx. est. n=5400). Also, only 36% of users' age and gender were known (est. n=15,000). At this point, permissible budget parameters were breached. In addition, GA will not show data for a specific group

until a threshold is achieved. There was a risk of the project running out of funds before data became accessible. As a result of the mounting costs and cost estimates, the research elected to pivot and follow a staged approach. A pilot stage of real time data was used to narrow the question set range 1-30 to a more affordable range. The exploratory stage used a narrower pool of randomised clustered groupings.

2.6 Research stages

The Pilot – In the pilot, the users were monitored using "Real-time" mode in GF and GA. Realtime mode allows a researcher to look at real-time data in a 30-minute window. The 30-minute window was monitored for 29 hours of the 48-hour period. The first goal was to confirm the random usage of each question set 1-30 by new users. Did the app insert code work? The second goal was to then monitor game level passing success of users as they returned to play the games. This required the use of both GF and GA; GF was used to monitor new users and GA was used to monitor the event of reaching and using a higher level of play. If a user was tagged for the first time opening of the app, then the number of questions per level was also tagged. This was logged and recorded. By looking at the window several times every hour, it was possible to build a picture of how random and uniformly spread the question sets 1-30 were. It was found that 29 of the 30 sets appeared within the first 68 recorded uses over a period of 17 hours. The final set was noted in the 107th recording, 12 hours into day two. The 107 live recorded users were from a total of 168 users that had joined during the full 48-hour period.

Exploratory stage – The app was reset. Then three further cohorts were created applying an iterative DBR approach. In cohort one, users were randomly given one of five question set lengths; 2, 6, 10, 15 or 19 (using the same random code approach as previously stated). These lengths were selected as representative of clusters based on the findings in Table 1. The data was clustered into sets 2-5, 6-9, 10-13, 14-17 and 18-21. One number was selected from each cluster. It was decided to stop at 19 because it had been demonstrated that approximately one in five return users had been willing to pass a 19+ question level quiz. It seemed likely that this number would overcome GA data thresholds. A brief analysis of the cohort one data led to a second cohort that were given lengths 1, 3, 7, 9 and 11. This then led to a third cohort provided question set lengths of 8, 12, 13 and 14.

Parabolic modelling – A quadratic regression predictive model was applied to the data. Essentially threading a path through 14 question-set length means, modelling a guiding estimate based on the data, of optimal and inhibiting selection.

2.7 Reliability and validity

Reliability was ensured by the automated design of the data collection. The same app was used throughout the experiment. The data was collected using the same mechanisms. In terms of validity, the measurements appear to have face validity. In fact, the instrumentation accurately measures the number of questions answered. However, in terms of sampling validity, non-probability convenience sampling nega-

tively impacts generalisability. This is discussed further in the discussion limitations section.

3 Findings

The focus of the research was on answering one key question. Does the number of questions in a quiz level impact total work completed? If yes, are there optimal and inhibiting question payloads delivered as a level set?

3.1 Initial results

The pilot study, appeared to show in broad terms, that quiz level does impact quantity of work completed. According to GF, of the recorded 168 users, 91 users returned to play more than one level, or one more session, during the 48-hour period. The research captured 73 of these levelling up user interactions in the 107, 30-minute window, sessions during researcher monitoring hours. These session clusters are very uniformly distributed. For example, the 107 recorded new users were distributed; 1-10 (n=37, 34%), 11-20 (n=37, 34%), 21-30 (n=33, 32%). When compared to the actual distribution of level passers, by length of question set, Table 4 suggests that a cluster of 1-10 question set users were almost 3.4 times as likely to return to the app and pass a level as the 21-30 question set cluster. This appears to be a very significant result. A chi square test of independence found there was a significant relationship between the three clusters and the distribution of level passers ($p < .001$) compared to the expected distribution of level passers. It would appear to support the hypothesis that the length of question sets does play a role in whether users progress beyond the first level. However, while a clarifying finding, this is not surprising. A small payload of two questions would allow many users to complete and be given the opportunity for a second payload. A large payload, such as 25 questions, was always likely to have less users complete and consequently many would not qualify for a second payload. However, we now have some data to support this view. Large question sets appear to inhibit usage.

Table 4. Clusters of ten lengths distribution of level passers

Question Sets	New User (n=107)	Level Passer (n=73)	Actual/Expected
1-10	34%	60.2%	1.77
11-20	34%	23.3%	0.69
21-30	32%	16.5%	0.52

chi square $p < .001$

The clustered results suggest the optimum number of questions per set are more likely to be in the lower question sets as users were more likely to return. However, more data was required to determine the optimum set size. But clearly higher question set lengths do seem to be inhibitors. The next step was to take a snapshot of a variety of question set lengths to see if a pattern would emerge.

3.2 Exploratory results

The exploratory results appear to confirm that quiz length both inhibits and optimises learning. A one-way analysis of variance, and a quadratic regression model, suggest that quiz level length has a parabolic relationship with learner output.

Long quiz lengths inhibit learning – An exploration of various question lengths strongly suggested optimal and inhibitor question length sets do exist. In the exploratory stage, the research looked at three cohorts. Each cohort was compiled to support or refute the findings in the previous stage or previous cohort. Cohort one considered five payloads, question sets: 2, 6, 10, 15, and 19. The focus was on users who were known to have started at least two levels and passed at least one level. The results in Table 5 are statistically significant ($p < .01$) and support the findings in the pilot. The install distribution was relatively even across the five payload lengths, but the distribution of those who levelled up, correlated to the number of questions in a set and mirrors the findings found in the pilot. Higher question lengths do seem associated with inhibitor effects.

Table 5. Cohort one distribution of level passers

Question Sets	New User (n=1017)	Level Passer (n=244)	Actual/Expected
2	20.5%	35.7%	1.74
6	18.8%	17.6%	0.93
10	20.5%	18%	0.88
15	21.2%	15.6%	0.73
19	19%	13.1%	0.69

chi square p = .005

Iterating towards optimum – The mean average answered questions, as shown in Table 6, provide a fairly accurate picture of which were generally the best quiz set lengths. There were two findings in cohort one that started the iterative narrowing optimisation process. As Table 6 cohort one shows, for motivated users question sets 10 and 15 are significantly more active but sets of 19 appear too large and six or less question lengths appear too small. The optimal, and inhibitor, effects are both starting to reveal themselves. To confirm the results of cohort one, cohort two was comprised of question set lengths 1, 3, 7, 9, and 11. It appears to follow a similar pattern. In this case, question lengths seven or less appear less than optimal, nine is of interest, but eleven is the optimal choice of the cohort. Please see Table 6 cohort two. The cohort three iteration was compiled based on the understanding that seven or lower and fifteen or higher appeared to be inhibitor options, while nine, ten, and eleven appeared interesting optimal possibilities. The data suggests (see Table 6 cohort three) that eight is the minimum size of interest, with a comparable result to a set of nine, while thirteen and fourteen are also interesting results, but twelve was the optimal choice. However, within cohort three there was no statistically significant difference between the four lengths.

Table 6. Cohort mean averages

Cohort One			Cohort Two			Cohort Three		
Set	***Mean***	***95% CI***	***Set***	***Mean***	***95% CI***	***Set***	***Mean***	***95% CI***
2	21.79	[17.00, 26.59]	1	24.28	[19.47, 29.09]	8	52.26	[43.30, 61.21]
6	35.44	[27.57, 43.32]	3	32.94	[25.54, 40.35]	12	70.24	[54.98, 85.49]
10	52.95	[39.49, 66.42]	7	37.06	[30.46, 43.67]	13	61.59	[48.30, 74.88]
15	46.18	[38.81, 53.56]	9	55.33	[43.98, 66.68]	14	53.89	[45.41, 62.36]
19	35.63	[29.41, 41.84]	11	61.98	[49.05, 74.90]			

One-way ANOVA ($p < .01$)

One-way ANOVA and Tukey post hoc test – A one-way ANOVA was performed on the raw user results of the question length sets to determine actual effect. Statistically significant findings were found ($F(13) = 8.85$, $p < .001$). However, a Tukey post hoc test revealed there were no, statistically significant pairwise differences between sets 1-7. There were also no, statistically significant pairwise differences between sets 8-15. But statistical significance was seen between these broad groupings and a third grouping comprised of set 19. The set length 12 appears to be the keystone. It reveals statistically significant pairwise difference to each set tested from 1-7, and to set 19 (see Figure 1).

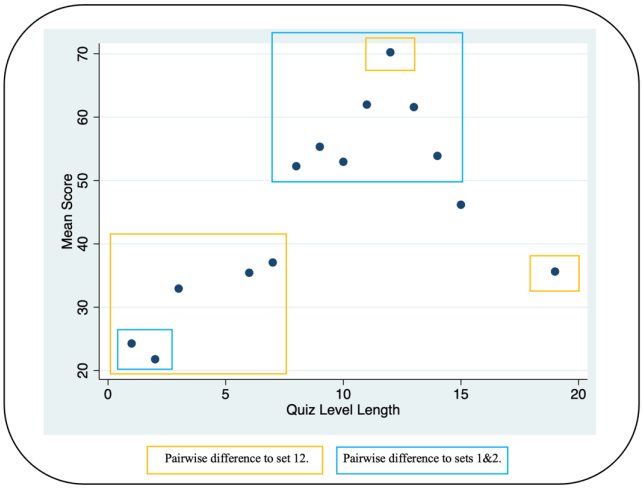

Fig. 1. Boxed significant pairwise difference

Furthermore, the mean average difference between set 12 and the listed sets, conservatively accounting for the 95% confidence intervals (see Table 6), was large: 25.9%-106.8%. As can be seen in Figure 2, this analysis was extremely conservative. While the data forms a peak at set 12, it is not definitive in defining set 12 as the absolute peak. What the data is suggesting is that there is a peak optimal zone and there are sub-optimal zones. There are three significantly different points that can be used to plot a parabolic graph. The zonal parabolic pattern can be identified in the boxed

areas of Figure 1. This is further supported by the Tukey ad hoc test results for set lengths one and two. Lengths one and two have no statistically significant difference to the other sets 1-7 nor 15-19 but are both statistically different to 8-14. This is approximately the inverse result of set 12, and we can deduce that they also support an optimal zone between set lengths 8-14.

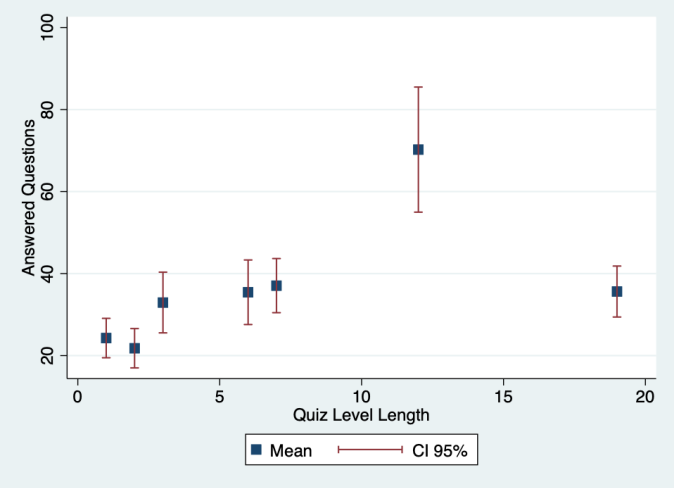

Fig. 2. Range of significant pairwise difference to set 12

Set one was effectively a control group. If a set one user answered one question they levelled up. There was no friction caused by question length. While each of the sets 8-14 could not be compared with each other, they could be compared to set one. It can be seen in Table 7 that sets 8-14 produce, at least, between 36% and 89% more mean answered questions than set one. Equally, again accounting for the 95% confidence intervals, sets 1-7, and set 19, produce only 0.48 to 0.79 of the mean answered questions that set 12 produces. In combination, we can deduce from these results that quiz length stimulates a quantitative work completed response and that response is likely parabolic, climbing to a peak and then declining.

Table 7. Significant level length comparisons

Set	Multiples of Set 1	Set	Multiples of Set 12
8	1.49	1	0.53
9	1.51	2	0.48
10	1.36	3	0.73
11	1.69	6	0.79
12	1.89	7	0.79
13	1.66		
14	1.56	19	0.76

Analysis conservatively factors in the 95% CI min/max range.

Quadratic regression predictive model – In order to further support the seemingly parabolic work completed response finding, a quadratic regression was performed to model the relationship seen between question set length and average mean of questions answered. A sample of 14 question lengths was used in the analysis. The results show that there is a statistically significant relationship between the explanatory variables question length and question length squared, and the response variable mean questions answered ($F(2, 11) = 19.07$, $p < .001$). Question length appears to account for 77.6% of variability in mean questions answered. The quadratic regression equation can be seen below and is visually represented in Figure 3.

Equation

$$\text{Predicted output} = 9.7125 + (7.9085 * \text{level length}) + (-.33655 * \text{level length}^2)$$

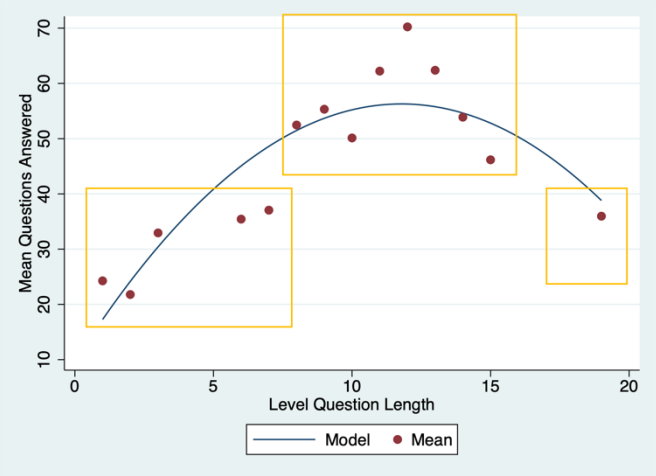

Fig. 3. Parabolic modelling and optimal zones

3.3 H_1 optimal payloads

The optimal zone **–** As stated, it is possible to construct optimal and suboptimal zones based on analysis of the pairwise statistical significance of set 12. Set 12 can be compared to sets 1, 2, 3, 6, 7, and 19 in the data, showing differences of 26%-107% after accounting for the 95% confidence interval. These strong results are supported in the literature by previously reported findings [40]. The results are statistically significant, and they show a large difference in user output. Using this comparative data, a border can be drawn, and sub-optimal regions can be defined. Sets 1-7 are suboptimal in comparison to set 12, as is set 19. Conversely sets one and two are statistically different to sets 8-14, while set 12 is not statistically different to sets 8-14. Again, the border is drawn, cutting seven from eight and 14 from 15. It can be stated that an optimal question set will be at least eight questions and tentatively less than 15, but 15 itself appears to be in a grey zone. Therefore, the optimal zone can be de-

duced as likely being within sets 8-14. But further research is required. See Figure 3 for a boxed outline of the broad optimal and inhibiting zones.

Peak payload not confirmed **–** The data suggests 11, 12, and 13 are all solid lengths and broadly support H_1. However, the data cannot verify which is truly optimal. All three appear reasonable choices. But, set 12, as already stated, can be compared to the largest number of suboptimal sets, and is further supported by the quadratic regression model. In addition, set 12 also has multiple factors that 11 and 13 do not possess. 12 is a particularly interesting number as it lends itself to further optimisation through component combination. The number 12 includes the factors two, three, four and six. This means we can include mini payloads within the question set to further optimise learning. For example, 12 questions, six new vocabulary items, four topics, three grammar points and two CEFR levels. It does seem likely that there are optimal pathways to learning and factor component combinations could be an interesting approach to materials optimisation.

3.4 H_2 inhibiting payloads

Larger payloads **–** There was evidence to suggest inhibiting payloads existed when quiz length became too long. The findings on inhibiting longer lengths were relatively unsurprising. It appeared users reacted increasingly negatively to quiz levels with over 15 questions.

Smaller payloads **–** The results were of genuine interest. The analysis suggests very small payloads such as two or three question sets did not lead to much work being done. They appeared to support H_2. They seem ultimately to inhibit learning potential and intention. They are not stretching people. The data suggests, see Table 8, over 40% of users may not have completed a three-question game level. However, if they do continue, they would be just as likely to complete eight questions. Two and three question sets are not efficient. The quadratic regression model further supports the notion that a low number of questions inhibits the amount of study completed. As can be seen in Figure 3, the lowest question set that provided reasonably strong results was a set of eight questions.

Table 8. Level one aborted

	1	**3**	**7**	**8**	**9**	**11**	**12**	**13**	**14**
Abort	0.163	0.429	0.466	0.445	0.542	0.577	0.601	0.593	0.669

4 Discussion

Analysis of the findings suggests quiz level length matters. The null hypothesis (H_0) can be rejected. Large differences in learner outcomes have been observed. The results appear to demonstrate the subconscious role of stimuli response within the operation, creating a statistically significant effect on the outcome of the activity. There is an optimal zone, however a meaningful singular optimal length cannot be definitively proven by the data, but it is likely to be at, or in the vicinity of, length 12.

The actual 26% gap between set seven and set twelve that is present in the conservatively adjusted data is large. It would be reasonable based on the quadratic regression model to see a gradual increment of approximately 5% between each added question from seven to twelve. However, an abrupt jump somewhere between seven and twelve cannot be discounted. A gradual increment would support a clean optimal peak while an abrupt shift would support a clean optimal zone. There is support in the data for an incremental hypothesis. Lengths one and two can be compared to 8-14, and length three can be compared to 11-13 and lengths six and seven can be compared only to 12. The narrowing is suggestive of incremental optimisation heading towards a peak at length 12. However, more data is required to concretely elicit such level of nuance. What can be stated is that the level lengths appear to coalesce within optimum and sub-optimum areas. There are clearly learning consequences to level length design decisions. These consequences can be seen in personal learning choice and the distortion of broader design decisions.

4.1 Personal choice

From the perspective of a self-study user, the problem is unintentionally limiting personal growth. Daily question apps are attractive marketable apps. The popularity of the apps is probably inspired by a mindset that a little bit of study is manageable or better than nothing. But if users understood the net result would be a lot less study, would it be as attractive? There is value in one question a day, if the alternative is zero study, but to a user who wants to pace themselves, they may have undermined their own learning goals. A conservative user would probably be better off using between 8-12 questions. EdTech developers should be more cognizant of the implications of their designs for their users' needs. The developers are probably not able to control or manipulate level length in all scenarios, but they could utilise it as part of personalisation packages. It is likely EdTech will need to educate both students and teachers, on the outcomes of level length, and allow the user to make optimal choices for their personal circumstances.

4.2 Game design distortion

Game design effectiveness could be distorted if level length is not controlled for and well understood. For example, if creating a gamified EFL quiz app, with the intention of promoting greater autonomous learning then it would seem question set length selection could be critical. The length could inhibit the gamification effect, or it could magnify it. A lack of understanding of how question length selection effects results, could in turn lead to over or under apportioning credit to any gamification strategy employed. For example, if we use the quadratic regression model data, which is more conservative than the raw data, look at Figure 3, selecting an inhibiting length, such as five would only produce 0.73 of the mean average of the optimal length 12. This is a very significant difference that must be considered. Put another way, switching from five questions to the use of 12 questions could magnify results

38% for a self-study approach. This is a large, conservatively modelled gain for a single simple adjustment.

4.3 Classical conditioning

The primary principle of classical conditioning is stimulus response. Level length variation stimulates a variation in response. The coupling of the completion of quiz questions and variable level length is leading to different quantities of work being undertaken. The optimal zone length covers a range of 8-14. Given base ten is how modern global society counts, it could be posited that the users are subconsciously expecting to complete the task in ten questions. Possibly a set of ten has the gravitas to hold user concentration and stretch them a little more. The app did not number the questions, the users received no cues as to how many questions they had answered. It is plausible that we have been conditioned by social norms to work approximately in tens, in much the same way as Pavlov's dogs [33], [34]. Furthermore, the lack of conscious self-awareness of this conditioning, means we can be stretched to answer between 11-14 questions before activating meaningful statistically significant resistance. However, further research is required.

4.4 Constructing new theory

Conceptually, activity length is a parabolic stimulator and constrainer of learning activity. The level of stimulation or constraint can be calculated using a parabolic prediction model based on quadratic regression. Consequently, variation in learner output can, to some extent, be predicted and optimised based on length of an activity. This is very well suited to the optimization of educational technology. Iterative application testing will allow for rapid output optimization based on activity length theory. It can also cross-apply to other professional environments. Although in this case, professional practitioners may benefit from the provision of pre-modelled optimization for various activity types. For example, for a multiple-choice quiz with one answer and two distractors, as found in this paper, the optimal model length is 12. This is based on the modelled data in Table 9 that was calculated using the quadratic regression equation as found in the findings.

Table 9. Modelled data for quiz level lengths 1-25

1	17.284	6	45.048	11	55.983	16	50.092	21	27.372
2	24.183	7	48.581	12	**56.151**	17	46.894	22	20.809
3	30.409	8	51.441	13	55.646	18	43.023	23	13.573
4	35.962	9	53.628	14	54.468	19	38.479	24	5.664
5	40.841	10	55.143	15	52.616	20	33.263	25	-2.919

4.5 A general model of activity duration

Arguably, for the theory to cross-apply to professional practice, it is likely that it must be developed into a more general theory applicable to all intentional activity in all contexts. It is highly likely that this will be a theory of activity duration. Time, will often, be the simplest measure for cross-application. As can be seen in Table 10, for this language learning app, one question can be modelled as taking approximately 20 seconds. An optimal time unit based on 12-questions is precisely four minutes (240 seconds). A 15-question level length, which appeared in the study to be of borderline significance, sets a limit for optimal activity at less than five-minutes (< 300 seconds). The minimum time unit would be greater than a length of seven (140 seconds), most likely about 2.5 minutes. In a classroom context, practitioners may find that creating action blocks (equivalent to game levels) of about 4 minutes within classroom tasks, leads to optimal output. They may also find action blocks greater then 5 minutes or less than 2.5 minutes are inhibiting output.

Table 10. Modelled activity duration of the optimal zone

Length	**1**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Seconds	20	140	160	180	200	220	**240**	260	280	300

This preliminary model data is tentatively provided as a starting point for cross-applying to professional practice. It is also likely that absolute precision is not required in many contexts and may vary between contexts. Therefore, it stands as a useful exploratory jumping off point, but more data is required to make the case for precise parameters.

4.6 Limitations

The non-probability sampling made the research possible but limits the generalisability of the findings. The research is at risk of under coverage bias, self-selection bias and non-response bias. This is partially mitigated to a certain extent by Thorndike's law of readiness [36] and the focus on self-study, self-motivated students; self-selection was a requirement of the research. The users had to want to do the activity. In addition, the collected demographic data suggests coverage has been relatively broad, and therefore the main issue is non-response bias. This likely bias does limit what can be extrapolated from the data. The research took place in a real-world global environment. The users faced many unknown impediments and were possibly receiving unknown rewards impacting their personal intention to study. Question set length was only one variable influencing study activity. It would be interesting to see the research at least partially replicated under experimental lab conditions. However, full replication is unlikely, the research allowed the participants to transfer intention to action over a nine-day period. This affordance is not likely under lab conditions, unless undertaken in a regimented environment, such as a prison or hospital. A further limitation is the sample size. While not small, it does not appear large enough to precisely account for all variation in the data. However, it is precise enough to show an

optimal zone. There are significant differences in mean average scores, uninstall rates and the most productive users, that combine to demonstrate a significant effect has occurred.

4.7 Future research

The construction of activity length theory and its application to various types of learning activity could be of great interest and could help professional practitioners in optimising student output and maximising user activity. The first step will be an examination of the influence of time on intentional activity. A postulated general theory of activity duration would then need to be tested with reference to different quantifiers of activity length, such as questions answered, utterances produced, or words written. The theory, it can be hypothesised, potentially applies to all activity; for example, games, listening activity, multiple choice quizzes, revision exercises, speaking activity, and timed writing. This will require extensive future research. However, the findings also lend themselves to other broader areas of concern. For example, behavioural research could be undertaken based on awareness of question set stimuli response. There could also be an argument for revisiting past research that may not have adequately factored in an activity length effect. In addition to m-learning application research, by cross-applying to professional environments, classroom-based studies could prove to be a rich vein of do-able research.

5 Conclusion

The major contribution of the paper is to bring into focus an aspect of mobile learning that was hidden in plain sight. Quiz level length significantly impacts total work completed. The null hypothesis H_0 has been rejected, and both H_1 and H_2 are supported. It is evident from the data, that learners have subconscious parameters for task completion, which may include a socially conditioned default to sets of ten. Furthermore, it is important to note, that providing learners with too little or too much work can lead to missed opportunities and negative responses, respectively. The optimal choice for quiz level length appears to be 12 questions on average, but this may vary from person to person and class to class. Awareness of the optimum set length is crucial; learners and teachers need individualised optimisation strategies. Multiple level length options seem appropriate for self-study technologies, with data driven formulations of what lengths might best suit the individual user. Additionally, it is recommended that the anticipated effect of length selection, on a user's learning, should be signalled to that user in advance. Self-awareness followed by a critical exploration of personal results will likely magnify individual ownership of learning routine. Learners will become more productive by understanding what quiz level length works for them. Moreover, the implications of quiz length choice may extend to professional environments. For example, classroom lesson effectiveness might also be impacted by teachers' quiz length choices. An awareness of the influence of activity length, could lead to a broader pedagogical look at how, in general terms, activity

length affects classroom learning. A theory of activity length could have significant implications for teaching, materials design and ultimately student outcomes.

6 References

[1] M. M. Elaish, N. A. Ghani, L. Shuib, and A. Al-Haiqi, 'Development of a Mobile Game Application to Boost Students' Motivation in Learning English Vocabulary', *IEEE Access*, vol. 7, pp. 13326–13337, 2019. https://doi.org/10.1109/ACCESS.2019.2891504

[2] C. Tejedor-Garcia, D. Escudero-Mancebo, V. Cardenoso-Payo, and C. Gonzalez-Ferreras, 'Using Challenges to Enhance a Learning Game for Pronunciation Training of English as a Second Language', *IEEE Access*, vol. 8, pp. 74250–74266, 2020. https://doi.org/10.1109/ACCESS.2020.2988406

[3] L. R. Octaberlina and I. Rofiki, 'Using Online Game for Indonesian EFL Learners to Enrich Vocabulary', *Int. J. Interact. Mob. Technol.*, vol. 15, no. 01, pp. 168-183, Jan. 2021. https://doi.org/10.3991/ijim.v15i01.17513

[4] A. H. Nabizadeh, J. Jorge, S. Gama, and D. Goncalves, 'How Do Students Behave in a Gamified Course?—A Ten-Year Study', *IEEE Access*, vol. 9, pp. 81008–81031, 2021. https://doi.org/10.1109/ACCESS.2021.3083238

[5] A. I. Zourmpakis, S. Papadakis, and M. Kalogiannakis, 'Education of preschool and elementary teachers on the use of adaptive gamification in science education', *IJTEL*, vol. 14, no. 1, pp. 1-16, 2022. https://doi.org/10.1504/IJTEL.2022.120556

[6] E. Dolzhich, S. Dmitrichenkova, and M. K. Ibrahim, 'Using M-Learning Technology in Teaching Foreign Languages: A Panacea during COVID-19 Pandemic Era', *Int. J. Interact. Mob. Technol.*, vol. 15, no. 15, pp. 20-34, Aug. 2021. https://doi.org/10.3991/ijim.v15i15.22895

[7] A. Maksum, E. N. Wahyuni, R. Aziz, S. Hadi, and D. Susanto, 'Parents' and children's paradoxical perceptions of online learning during the Covid-19 pandemic', *Adv Mobile Learn Educ Res*, vol. 2, no. 2, pp. 321–332, 2022. https://doi.org/10.25082/AMLER.2022.02.002

[8] K. Lavidas, Z. Apostolou, and S. Papadakis, 'Challenges and Opportunities of Mathematics in Digital Times: Preschool Teachers' Views', *Education Sciences*, vol. 12, no. 7, p. 459, Jul. 2022. https://doi.org/10.3390/educsci12070459

[9] S. Papadakis, A. İ. C. Gözüm, M. Kalogiannakis, and A. Kandır, 'A Comparison of Turkish and Greek Parental Mediation Strategies for Digital Games for Children During the COVID-19 Pandemic', in *STEM, Robotics, Mobile Apps in Early Childhood and Primary Education*, S. Papadakis and M. Kalogiannakis, Eds. Singapore: Springer Nature Singapore, 2022, pp. 555–588. https://doi.org/10.1007/978-981-19-0568-1_23

[10] A. Tzavara, K. Lavidas, V. Komis, A. Misirli, T. Karalis, and S. Papadakis, 'Using Personal Learning Environments before, during and after the Pandemic: The Case of "e-Me"', *Education Sciences*, vol. 13, no. 1, p. 87, Jan. 2023. https://doi.org/10.3390/educsci1301–0087

[11] S. Papadakis, F. Alexandraki, and N. Zaranis, 'Greek Parents' App Choices and Young Children's Smart Mobile Usage at Home', in *New Realities, Mobile Systems and Applications*, vol. 411, M. E. Auer and T. Tsiatsos, Eds. Cham: Springer International Publishing, 2022, pp. 39–50. https://doi.org/10.1007/978-3-030-96296-8_4

[12] J. Byrne, 'Same time same place: Do MALL classrooms exist?', *Teaching English with Technology*, vol. 16, no. 3, pp. 74–84, 2016.

[13] J. S. Lee, 'Informal digital learning of English and second language vocabulary outcomes: Can quantity conquer quality?: Informal digital learning of English', *Br J Educ Technol*, vol. 50, no. 2, pp. 767–778, Mar. 2019. https://doi.org/10.1111/bjet.12599
[14] J. Byrne, 'Anytime Autonomous English MALL App Engagement', *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 18, pp. 145–163, Sep. 2019. https://doi.org/10.3991/ijet.v14i18.10763
[15] R. Fithriani, 'The Utilization of mobile-assisted gamification for vocabulary learning: Its efficacy and perceived benefits.', *CALL-EJ*, vol. 22, no. 3, pp. 146–163, 2021.
[16] M. Nurtanto, N. Kholifah, E. Ahdhianto, A. Samsudin, and F. D. Isnantyo, 'A Review of Gamification Impact on Student Behavioural and Learning Outcomes', *Int. J. Interact. Mob. Technol.*, vol. 15, no. 21, pp. 22-36, Nov. 2021. https://doi.org/10.3991/ijim.v15i21.24381
[17] B. Waluyo and J. L. Bucol, 'The impact of gamified vocabulary learning using Quizlet on low-proficiency students', *CALL-EJ*, vol. 22, no. 1, pp. 164–185, 2021.
[18] P. Pando Cerra, H. Fernández Álvarez, B. Busto Parra, and P. Iglesias Cordera, 'Effects of Using Game-Based Learning to Improve the Academic Performance and Motivation in Engineering Studies', *Journal of Educational Computing Research*, pp. 1663-1687, Jan. 2022. https://doi.org/10.1177/07356331221074022
[19] F. Ugur-Erdogmus and R. Çakır, 'Effect of Gamified Mobile Applications and the Role of Player Types on the Achievement of Students', *Journal of Educational Computing Research*, pp. 1063–1080, Jan. 2022. https://doi.org/10.1177/07356331211065679
[20] K. M. Kapp, The gamification of learning and instruction: game-based methods and strategies for training and education. San Francisco, CA: Pfeiffer, 2012. https://doi.org/10.1145/2207270.2211316
[21] J. McGonigal, *Reality is broken: why games make us better and how they can change the world*, Ed. with a new appendix 2. New York: Penguin Books, 2011.
[22] B. Burke, *Gamify: how gamification motivates people to do extraordinary things*. Brookline, MA: Bibliomotion, books + media, 2014.
[23] D. Hathaway and P. Norton, *Understanding Problems of Practice*. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-77559-3
[24] R. Huang, J. M. Spector, and J. Yang, 'Design-Based Research', in *Educational Technology*, Singapore: Springer Singapore, 2019, pp. 179–188. https://doi.org/10.1007/978-981-13-6643-7_11
[25] J. Hardman, 'Researching pedagogy: An activity theory approach', *Journal of Education*, vol. 45, no. 1, pp. 65–95, 2008.
[26] A. Sannino and Y. Engeström, 'Cultural-historical activity theory: founding insights and new challenges', *Cultural-Historical Psychology*, vol. 14, no. 3, pp. 43–56, 2018. https://doi.org/10.17759/chp.2018140304
[27] L. S. Vygotsky, *Thought and language*, 17. print. Cambridge, Mass: MIT Press, 1985.
[28] A. N. Leont′ev, *Activity, consciousness, and personality*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.
[29] T. D. Wilson, 'A re-examination of information seeking behaviour in the context of activity theory', *Inf. Res.*, vol. 11, 2006.
[30] Y. Engeström, *Learning by expanding: an activity-theoretical approach to developmental research*, Second edition. New York, NY: Cambridge University Press, 2015. https://doi.org/10.1017/CBO9781139814744
[31] E. L. Thorndike, 'The Law of Effect', *The American Journal of Psychology*, vol. 39, no. 1/4, pp. 212-222, Dec. 1927. https://doi.org/10.2307/1415413

[32] B. F. Skinner, 'The Generic Nature of the Concepts of Stimulus and Response', *The Journal of General Psychology*, vol. 12, no. 1, pp. 40–65, Jan. 1935. https://doi.org/10.1080/00221309.1935.9920087

[33] I. P. Pavlov, 'Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex', *ANS*, vol. 17, no. 3, Jun. 2010. https://doi.org/10.5214/ans.0972-7531.1017309

[34] O. Bichler *et al.*, 'Pavlov's Dog Associative Learning Demonstrated on Synaptic-Like Organic Transistors', *Neural Computation*, vol. 25, no. 2, pp. 549–566, Feb. 2013. https://doi.org/10.1162/NECO_a_00377

[35] A. Budiman, 'Behaviorism and Foreign Language Teaching Methodology', *Engl. Franca acad. j. of Engl. lang. and educ.*, vol. 1, no. 2, pp. 101-114, Dec. 2017. https://doi.org/10.29240/ef.v1i2.171

[36] Y. Ni and J. Lu, 'Research on Junior High School English Reading Class Based on the Principle of Timing and Thorndike's Three Laws of Learning', *JLTR*, vol. 11, no. 6, pp. 962-969, Nov. 2020. https://doi.org/10.17507/jltr.1106.13

[37] T. Koyama and O. Takeuchi, 'Does Look-up Frequency Help Reading Comprehension of EFL Learners? Two Empirical Studies of Electronic Dictionaries', *CALICO Journal*, vol. 25, no. 1, pp. 110–125, 2007. https://doi.org/10.1558/cj.v25i1.110-125

[38] C. T. X. Lien and L. T. H. Phuong, 'Using Moodle Quiz to Assist Listening Assessment: EFL Students' Perceptions and Suggestions', *Journal of Inquiry into Languages and Cultures*, vol. 4, no. 1, 2020.

[39] Nguyen Van Bao and Nguyen Van Loi, 'MOODLE QUIZ TO SUPPORT VOCABULARY RETENTION IN EFL TEACHING AND LEARNING', Mar. 2020, doi: 10.5281/ZENODO.3708413.

[40] A. Kobren, C. H. Tan, P. Ipeirotis, and E. Gabrilovich, 'Getting More for Less: Optimized Crowdsourcing with Dynamic Tasks and Goals', in *Proceedings of the 24th International Conference on World Wide Web*, Florence Italy, May 2015, pp. 592–602. https://doi.org/10.1145/2736277.2741681

[41] J. Byrne, 'Southeast Asian Short-Burst Parameters for Autonomous Mobile Learning: One Step toward Automated Situated MALL', *Ubiquitous Learning: An International Journal*, vol. 13, no. 2, pp. 31–42, 2020. https://doi.org/10.18848/1835-9795/CGP/v13i02/31-42

[42] Unity Technologies, 'Unity'. 2021. Accessed: Nov. 01, 2021. [Online]. Available: https://www.unity3d.com

[43] Mocapot, 'Ultimate trivia quiz game kit'. Mocapot Game Studio, Mar. 08, 2021. Accessed: Nov. 05, 2021. [Online]. Available: https://assetstore.unity.com/packages/templates/packs/ultimate-trivia-quiz-game-kit-174318

[44] 'Firebase Analytics'. Google, Mountain View, CA. Accessed: Feb. 15, 2023. [Online]. Available: https://firebase.google.com/docs/analytics

[45] 'Google Analytics'. Google, Mountain View, CA. Accessed: Feb. 15, 2023. [Online]. Available: https://marketingplatform.google.com/about/analytics/

[46] M. M. H. J. van Gelder, T. H. van de Belt, L. J. L. P. G. Engelen, R. Hooijer, S. J. H. Bredie, and N. Roeleveld, 'Google AdWords and Facebook Ads for Recruitment of Pregnant Women into a Prospective Cohort Study with Long-Term Follow-Up', *Matern Child Health J*, vol. 23, no. 10, pp. 1285–1291, Oct. 2019. https://doi.org/10.1007/s10995-019-02797-2

[47] U. D. Upadhyay, I. J. Jovel, K. D. McCuaig, and A. F. Cartwright, 'Using Google Ads to recruit and retain a cohort considering abortion in the United States', *Contraception: X*, vol. 2, p. 100017, 2020. https://doi.org/10.1016/j.conx.2019.100017

[48] World Medical Association., 'World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects', *Bull World Health Organ*, vol. 79, no. 4, pp. 373–374, 2001.
[49] General Assembly of the World Medical Association, 'World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects', *J Am Coll Dent*, vol. 81, no. 3, pp. 14–18, 2014.
[50] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research., 'The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.', The Commission, Bethesda, Md, 1978.
[51] HHS, '45 CFR 46', *Exemptions (2018 Requirements)*. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/common-rule-subpart-a-46104/index.html (accessed Feb. 14, 2022).

7 Author

Jason Byrne is an Associate Professor at INIAD, Toyo University, Tokyo 115-8650, Japan. Jason is also, as of the time of publication, affiliated with Tokyo Denki University. Byrne has co-authored multiple 1 million download English study apps. His interests include CALL, digital nudging, gamification, and m-learning (Email: byrne@toyo.jp).

Article submitted 2023-01-28. Resubmitted 2023-03-01. Final acceptance 2023-03-11. Final version published as submitted by the author.