Digitally Nudged Learning:

A Nudged Gamification Study Intervention

https://doi.org/10.3991/ijet.v17i12.30567

Jason Byrne(✉), Takehiko Ito, Mariko Furuyabu
Toyo University, Tokyo, Japan
byrne@toyo.jp

Abstract—The research focused on two elements of technology enhanced learning: the digital nudge and gamified online learning. The objective was to digitally nudge students towards a gamified online learning tool, thereby improving quiz test performance through a fun motivating language learning game. A mixed methods approach: the primary quasi-experimental methodology was regression discontinuity design, with a follow up survey. The findings show that few students actively participated in the fun gamified activities, yet none-the-less there was a significant 5% treatment effect afforded by the digitally nudged call to action. The research demonstrates the effectiveness of nudging students to complete L2 learning activity. It also somewhat shows the potential of gamification for voluntary engagement amongst first-year students at a Japanese university.

Keywords—digital nudge, online learning, gamification, l2 teaching/learning strategies

1 Introduction

The research looked at a potential study intervention to improve vocabulary quiz scores within one university faculty. The focus was on the lower performing students among a cohort of first year university students, offering the students a gamified online treatment option. The research questioned whether a teaching strategy of simply nudging students with an offer of gamified study would improve scores. The research also questioned how likely students are to take up the treatment offer, and if they did, what was the impact on posttest results. It was hoped that an evaluation of the research results may lead to improvement in effective teaching strategies within the faculty.

Gamification is a relatively new, 21st century, approach to motivating a user to act, coming to prominence around 2010 [1]. It has been applied to language learning, for example, see [2]-[6], and can be found in the design principles of mobile assisted language learning (MALL) apps. Some examples of gamified apps known to have been used in L2 research are Duolingo [7], Kahoot [8], and Quizlet [9]. Gamification relies on simple principles; purpose, rules, voluntary participation, an obstacle to overcome, and a feedback system [10]. Furthermore, games provoke emotional responses; a sense

of fiero (pride) when things go well and ideally a sense of happy failure when things do not go so well [10]. Together this creates an experience that motivates participants to play more and reach their goals. This fundamental approach can be used for education. Games can be both motivational and instructional, and in combination are a particularly powerful tool [11]. So powerful, that users have been documented playing literally 24/7 [12]. In the case of this English L2 learning research project, the goal of student participation was to improve vocabulary quiz scores. The rules of the game were provided by the selected gamified application. The primary obstacle was acing the game-based quizzes, while secondary obstacles took the form of in-game anonymized classmate competition. Feedback was instantly provided after each question attempt. The missing piece, was would the students voluntarily participate? It is stated in [13], and this cannot be re-stated enough, gamification is about helping users to reach their own goals. It cannot be only the teacher's goal or the school's goal; this will not work. The student must be immersed voluntarily into the experience. If a student wants to improve, then gamification is a method to leverage that ambition and drive that student forwards.

The treatment involved offering students gamified language learning activity. But the act of offering is itself a call to action. According to nudge theory [14], we, the research practitioners, took on the role of choice architects. There was no coercion or demand for students to change, instead a choice was provided that the students were free to ignore. Digital nudging [15], is the use of digital design elements to influence behaviour in digital choice environments. It is particularly important as education moves increasingly into digital choice environments, such as Google Classroom [16] and MALL. For example, nudging users via in-app notification is a common feature of many apps. It is done because it is known to increase participation and it is constantly being improved, for example, see [17]. Nudging can be more subtle than the aptly named push notification, for example a flashing button, an alarm sound or colour guided pathway. We are now seeing the emergence of AI nudging that will exponentially impact all areas of a nudged life [18]. Recent research [19], shows nudging does have an effect in educational settings. Therefore, it is likely that even if students did not accept the offer to play a gamified app, it would impact their study choices. They would possibly be more mindful of their up-coming class vocabulary quiz. It may have prompted them to study a little harder. It was anticipated that this would be reflected in the treatment effect. However, while there are clear benefits, there are also concerns with nudging people to change behaviours [20]. It raises legitimate questions about autonomy and manipulation. However, fundamentally teaching is concerned with helping students to improve themselves. The goal of teaching is always to affect change. In this case test scores. That said, students should always have the option to opt out and this was reflected in the hypothesis and approach to this project.

1.1 Hypothesis

Opt-in digitally nudged gamification can be used as a teaching strategy to increase student vocabulary test scores.

The research question from a practitioner perspective attempted to see if we could guide students towards more interesting materials, and if in doing so, we could help improve their learning outcomes. Borrowing Thaler & Sunstein's terminology [21], the experiment can be viewed as pedagogical libertarian paternalism. The objective is to kindly help students improve, but while keeping mindful of the fact that they were free to decline the said kind offer.

2 Methods

A mixed methods approach, two methods were used in the study. Initially, regression discontinuity design was implemented with a pretest selection rating, treatment, and posttest. This was then followed up with a small survey of the treatment group.

2.1 Participants

The research took place at a university in Tokyo, Japan. The students were 18-22 years of age with only one student outside this range. The faculty student population are predominantly men with a gender ratio of 4.23:1.0 (see Table 1). Due to the limited size of the population (N=392), the research did not segment the data by gender nor age. In addition, foreign national non-resident students coming from overseas were not included in the data as they do not take compulsory English classes. The focus of the study was on treatment versus control. Simply, did the treatment influence first-year student compulsory English class results?

Table 1. First year demographics

Age		**Gender**	
18	49	Men	317
19	247	Women	75
20	82		
21	12		
21>	2		

The research initially focused on all first-year students taking a compulsory reading and writing class. The students were required to pass the class to graduate. The course had been taken online by all first years, effectively in distance education mode. In 2021 this was the new normal in COVID-19 Japan. All students received the same materials with the same teacher led videos. There was also an actual class teacher available for contact and the students were in class groups of 40, many of whom they knew from on-campus Listening and Speaking classes. The test phase took place over a two-month period. Different vocabulary items and questions were used on the pretest and posttest. Each quiz comprised 10 questions with each question valued at one point. The faculty has ten reading and writing classes; two classes were removed from this study in order

to improve participant anonymity. In addition, before the treatment group selection began, further students were deselected for missing quizzes. This also improved anonymity. The cohort of 271 remaining students were members of eight classes with four different teachers, the rating and posttest scoring between the eight classes was identical, all class tests used a Google form, and the quizzes were marked automatically. For clarification, the faculty had been using all the various online learning cloud-based solutions mentioned for five years, the only difference with COVID-19, was that this specific course had also become video lecture distance education.

2.2 Regression discontinuity design

It was decided to use retrospective regression discontinuity design (RD) in preference to randomized experimental design. RDs are increasingly used in educational settings to obtain estimates of intervention effects [22], for example, see [23]-[25]. Regression discontinuity design (RD) is a quasi-experimental approach that is particularly suitable for education [26]. RD can be undertaken retrospectively but requires additional graphical and empirical analysis [27]. RD is often used to determine merit or need [28]. In this study the focus was on the need to improve student vocabulary scores. The basic principle of RD is to pretest participants, and then position them by rating on a linear line with a predetermined cut-off point for intervention. A simple linear, or local polynomial, regression analysis of the data, collected after the applied intervention, it is generally anticipated, will, if a positive effect exists, lead to an upward shift in the line at the cut-off point, creating two separate lines. The size of the shift as seen on a graph, visually encapsulates the effect.

2.3 Ethics

There are ethical considerations when undertaking research into student performance. Should practitioner researchers intentionally not help underperforming students for the sake of experimental procedure? There is a genuine ethical dilemma in withholding treatment from suitable candidates [29]. RD offers an ethical alternative to experimental classroom research, as the researchers can focus treatment on an identified need immediately. However, the pros and cons of random control versus regression discontinuity design must be carefully weighed. In this case, it was a simple decision. There was little to be gained by randomly deciding not to help students. It made sense to prioritise helping students now and giving as many as possible a nudged gamification experience. Furthermore, as previously stated, the core intention of the research was to evaluate this potentially innovative teaching strategy with a view to using it with all students in the near future. It made sense to see how it would help those who most need it.

Human research. In terms of ethical human research there were two aspects to consider: score data and treatment participation. The first aspect is the use of low-level pretest and posttest educational score data. This is generally exempted from human research policies if it will not have adverse effects on the students. For example, the US federal policy 45 CFR 46.104-d-1 [30]. This exemption includes, educational score

data, and research into instructional techniques and strategies. All data collected (pretest and posttest) that identified students, fell into this category of data. However, confidentiality and privacy were respected. The data was anonymized as soon as was possible, there was no analysis of identifiable individual participants, beyond the initial selection of the treatment group.

Informed consent for treatment participation. The students were sent an email, in Japanese, explaining that a treatment was available to help them improve their quiz grades for the latter half of the semester. It was stated that the treatment was part of a research project and student participation was voluntary. It was further explained that the activities were anonymous and non-participation would have no adverse consequences. It was also advised that they had been selected as they appeared to need the treatment and could benefit from participating. Since the activity was distance and online, the actual consent was construed to have been given if the student followed the two-step process and opted-in to the treatment Google Classroom. Furthermore, no individualized data was collected on the specific treatment activity. This means we did not monitor what individuals did during the treatment. Students were informed of this fact.

2.4 Validity

During the research, attention was placed on the ten most general threats to validity [31], and two significant potential threats were noted, statistical regression and research mortality. Both are covered in the broader topic of RD Design validity below. There are unique considerations to ensure the validity of RD results. A sample size multiple must be applied to RD to see effects comparable to a randomized trial; in [32] it is stated that a sample size multiple of 2.75 is required for a balanced design, normal distribution rating. In educational research typically 30 participants are the minimum sample size, and 15 participants should be the minimum group size. This suggests a minimum of 84 participants are required as a treatment group of 42 would be preferred for educational RD. The research met the standard and initially had a treatment group n=128 and control n=143, but this must be revisited as attrition took its toll.

For the further purposes of RD validity, participants' ratings and the cut point must be determined independently of each other [32]. In real world terms, ideally the cut point would not be known to the teachers or students. It can be confirmed that the students had no knowledge of the research project or treatment at this early stage. It was anticipated a score of seven out of ten for each of the three quizzes which equates to 21 points or 22 points (rounded down), would be a potential cut-off point to create a large enough sample for inclusion in both treatment and control groups. It was generally understood by the researchers and teachers that inclusion would mean only a 50% chance of being offered the treatment and presumably decreased teacher interest (if any) in manipulating results. In addition, it was the non-teaching researcher who suggested RD. Furthermore, the rating used for the RD design was a series of quizzes that partially accounted for students' grades. The pretest was administered by Google Form and was taken at home, due to COVID-19 restrictions. The students had one chance to take the test. The results were immediately known to the student and with one click transferred

to Google Classroom for classroom teacher inspection. There was very little chance for a teacher to subconsciously favour student selection and manipulate the results on who would be placed in the research sample and even less motivation.

2.5 RD procedure

RD design follows a staged process and can require multiple iterative procedural steps: research design creation, a visual data check, treatment, posttest, a final attrition check, followed by sensitivity testing. These steps are outlined below.

RD creation (step 1). A study requires three basic components to be considered RD [22]. There must be a forcing variable, with participants placed into treatment and control groups based on being either higher or lower than the forcing variable. Secondly, the forcing variable must be ordinal (order the participants) and must include at least four values above and below the cut-off point. Thirdly, the cut-off value must not be used for assigning participants to other interventions. These three criteria were met; a pretest score was used as the forcing variable, with five values above and twenty-five values below the cut point. It was not being used for any other intervention.

The rating pretest materials, comprised of a series of three vocabulary quizzes valued at a total of 30 points, were delivered to and subsequently completed by 271 students. These vocabulary tests were taken at home, due to COVID-19. It was anticipated that most students would do very well. They could check dictionaries, undertake collocation searches, so there was no reason for them not to score highly.

Visual examination of the data (step 2). The pretest ranked the 271 students on a scale from 0-30. All scores would fall precisely at points along a linear line on a graph. This is ideal for RD as there is clear guaranteed linearity at the pretest stage, prior to the application of linear regression at the posttest stage. As can be seen in Figure 1, the data is not actually scattered, but clustered on 17 rating points, it follows a clear linear path.

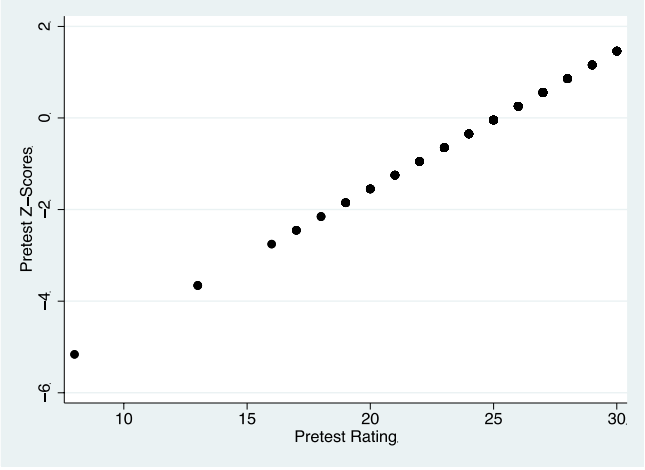

Fig. 1. Pretest scatterplot

It was found that the initial pretest results formed an approximately normal distribution with 74.5% of students falling within one standard deviation of the mean, with approximately 12.6% having a standard z-score of < -1 and 12.9% with z-scores of > +1. The standard deviation of the rating data is 3.2 points with a mean score of 25.15; 95% CI [24.95, 25.35]. However, it was not a standard normal distribution, given the mean, median and mode were not equal. The mean score was 25.15, the median was 26 and the mode jointly 26 and 27 points. [33] states that 50% of the scores should fall within the range 0.67 of the standard deviation from the mean for a standard normal distribution. In this case, 158 from 271 scores, 58.3% fall within the range. Furthermore, this slightly negatively skewed result is reflected in the fact there is a poor performing tail. The poor performers were spread more widely than the more closely clustered strong performers. This can be seen in the raw pretest score point distribution (please see Figure 2).

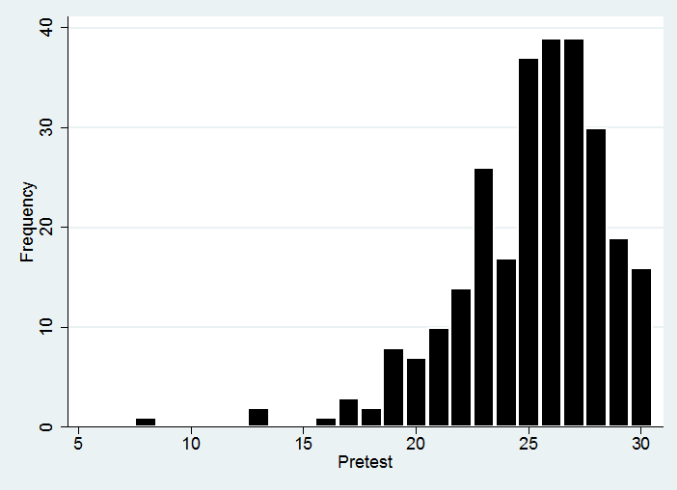

Fig. 2. Distribution frequency of pretest scores

The originally considered cut-off point, 22 points, was approximately one standard deviation from the mean. Therefore, a poor choice, due to validity issues with statistical regression and sample size. The 34 students who had z-scores below -1 (8-21 points) are, as can be seen in Figure 2, performance outliers and quite likely to improve without treatment and will, if too large a proportion of the sample, distort the results. 22 points itself (n=14), provided too small a sample to counter the effects of the outliers or to be studied in isolation. Using the mean of 25.15 or the median of 26 points as the cut-off point were the only remaining options that made sense based on empirical analysis of the data. The median is the optimum choice as it places in theory 50% of students in either group. In fact, frequency impacted the selection. 26 points were scored by 39 students. This meant if included in the treatment, then the treatment was 62% of the sample. If not included, then the treatment invitees were 47% of the sample. Given the results were close to the extreme in terms of needing a minimum of four rating values

above the cut-off point, it was decided that the cut-off point should be below 26 points, in effect z, the mean of 25.15. All scores were converted to z-scores. All students with standard scores below z = 0 were in the treatment and all above z = 0 were in the control. Since no student had a standard score of precisely zero there was no ambiguity. It was concluded based on the above graphical and empirical analyses that the design is valid and the decision to use a z-score of zero (mean = 25.15 points) as the cut point, allowed the practitioner researchers to help the below average students (47%) who had taken all three pretest quizzes.

RD treatment and posttest (step 3). A two-step, digitally informed-consent, invitation process was undertaken. Students were invited (nudged) by email to the treatment group to undertake extra gamified study. It was explained that data would be collected for research purposes. It was further explained that all activity (or lack of) on Google Classroom and Kahoot would be anonymous. The invitations were sent a maximum of three times as required. They were invited to a Google Classroom made specifically for the activity. If they joined the Google Classroom, then they had opted-in to the study. The opted-in students were provided with three mini (five question) practice games made using Kahoot for each of the three posttest quizzes. The mini-practice games were simple quiz format questions. Kahoot is a gamified quiz platform, providing randomisation with game elements such as sound effects and leaderboards. Users are anonymous by default. It is also possible to engage in low stakes anonymous competition with classmates, for example, see [8].

The opted-in students were then nudged, twice a week via email, to practice on Kahoot, in the lead up to the actual quiz which was provided as a Google Form. In total the students were nudged six times for the three posttest quizzes. The nudge was an email, sent with intention using the course leader's email address. It was anticipated that seeing an email from the course leader would be read, as opposed to Google Classroom automated emails which the students would receive from courses on a regular basis and whose effect may have been diluted. The email message was short and simple, with a direct link to the Google Classroom material for that week. It was automatically scheduled to be sent at 5pm on Wednesdays and Saturdays prior to the actual quiz deadline of the following Tuesday.

Attrition & design check (step 4). The study sample size was initially 271 participants, but this fell to 212 participants, control n=121 and treatment group n=91 after removing no shows. The no shows included 19 participants invited to the treatment who missed one or more of the posttest quizzes and a further 18 who did not opt-in to the study. Consequently, the opt-in treatment group had 91 members. The control group was also decreased by 22 participants to 121 members for missing one or more of the posttest quizzes. In terms of the impact of experimental mortality, 14.8% of the treatment invitees and 15.4% of the control were deselected for missing quizzes. Experimental mortality affected both treatment and control to a very similar degree. Furthermore, the groupings still met the minimum size validity condition set out in step 1. Importantly, in addition, the final cohort (n=212) had a retrospective pretest mean of 25.3. Therefore, the experimental cut-off point, set to z = 0 in step 2, did not significantly change as a consequence of the research attrition. A fuzzy RD design is required when the number of no-show participants leads to distortion [28], that effectively means

active participants have been misassigned to the wrong grouping (control or treatment). In this experiment no one was misassigned, as while the position of z had shifted slightly from the old mean of 25.15 to the new mean of 25.3, it made no difference to group assignment. All students with 25 points or less, would still be selected for treatment and all those with 26 points or more, would still be in the control. Nothing had changed, and so based on this understanding of the data, a sharp RD design was selected for use. In fact, there was a change, the data was now approximately normally distributed. 70.7% of students fell within one standard deviation (SD), with 17% below -1 SD and 12.3% at +1 SD. The tail to the left reflected the fact the majority of students were scoring highly, and this shifted the centre of the distribution rightwards on the graph.

To check for reliability, an RD density test [34], was undertaken, using Stata software [35], and did not reject (p=.22) the null hypothesis of no manipulation. In other words, no falsification of the data was found.

Graphical analysis (step 5). A scatterplot of the posttest outcome against the pretest rating is shown in Figure 3. This graph shows extensive movement encapsulating the expected regression to the mean and potentially a treatment effect. Initially, the groups were split at zero based on pretest data. Approximately 40% of students rated in the z-score range -2 to 0 have moved upwards into the 0 to 2 zone, with around 10% moving downwards into the lower outlier areas. Similarly for the control students rated with z-scores of 0 to 2, visually we can see that their outcomes have decreased. Approximately 50% have dropped into the -2 to 0 area. This has created an almost rectangular mesh pattern in the top right quadrant of Figure 3, where in Figure 1, after the initial pretest, there had been a clean straight line.

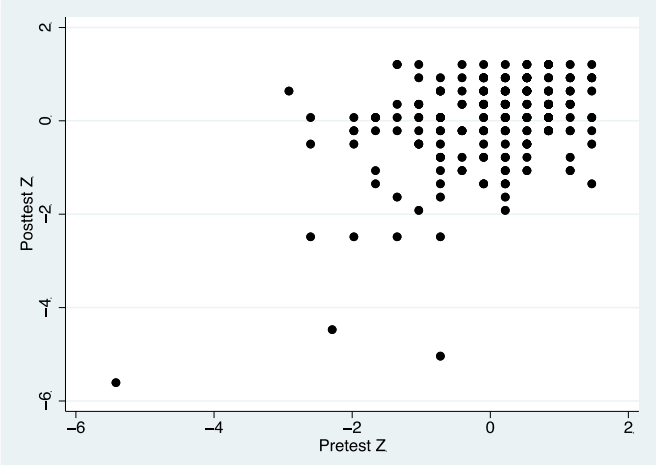

Fig. 3. Scatterplot of outcome

Functional form (step 6). The selection of functional form is of critical importance, as it makes sure that the chosen form is not itself adding bias to the estimate. The first procedure was to decide whether to use a parametric or nonparametric form. Given the

relatively small scale of the project, it is said a parametric global strategy is the appropriate choice [27]. However, in [36] it is stated that in recent RD design, non-parametric local polynomial approximation is considered the preferred choice. Bin sensitivity testing led to a clear conclusion that a nonparametric local polynomial estimation was the correct form. Given, the project had standardized the data, a perfect score of 30 points equates to z = 1.46. The primary idea of RD Design is to split the bell curve and analyse the mirror image as treatment effect and regression shift the line. However, 24 from 91 treatment participants ranged in z-scores from -1.5 to -5.4. They had no positive mirror image to counter and consequently were seriously distorting, almost nullifying, the results of treatment effect among typical students closer to the cut-off point, irrespective of bin size. As an initial step the bandwidth was set at a z-score of = ±1.5 before it was finally set to z = ±1.0.

To minimize noise, data bins were created. Given the small sample size this was a challenge, and many options were tried: 4-bin, 6-bin, 7-bin, 8-bin, 9-bin, and 10-bin combinations. The process was greatly simplified once the noise created by the outliers was removed. The concrete choice was to use the original pretest rating points as bin indicators. At bandwidth z = ±1.5 there were five ratings on either side of the cut, meaning 10 bins: five treatment and five control. This naturally falls to the second option of six bins at bandwidth z = ±1. There were three other reasonable options: 6-bin sets based on standard deviation indicators at bandwidth z = ±1.5, a simple split into four bins, or the fourth option was to use the raw data from the localized sample and not apply bins. At this stage it was an iterative process of test and eliminate. Firstly, the 4-bin split can only be used with the linear model, and the result was found to be not significant (p-value of the F-test =.11). The 4-bin approach was eliminated. Furthermore, the z = ±1.5 bandwidth produced inconsistent results, the linear results show almost no effect while the quadratic showed a strong effect, this suggests the data still included outlier distortion. The bandwidth z = ±1.5 option was eliminated. The selection of bandwidth is a somewhat subjective process. A computed bandwidth check, using the rdbwselect function of the rdrobust package [37], selected z = ±0.992. In effect this is z = ±1, as there are no values between the two. The data for z = ±1 behaved, but there were now only two options, as the 10-bin model had folded to a 9-bin significant result at z = ±1.5 and a 6-bin at z = ±1 (See Table 2).

Table 2. Bandwidths vs bin size F-test p-values

Bandwidth/Bin Size	Raw Data	10-Bin	9-Bin	6-bin
z = 1.5	0.23	-	0.035	0.12
z = 1.0	0	-	-	0.014

By changing the bandwidth to z = ±1 there were now two choices: six bins or raw data, and both, as Table 2 shows, were statistically significant ($p<.05$). The raw data option during coefficient computation would also rely on the same six values of x that the 6-bin option comprised. Consequently, it would be expected, if there was any underlying consistency, that the raw data and 6-bin coefficients would be very similar. In fact, as Table 3 shows, the linear models are within 5% of approximation, but the quadratic model has shifted the outcome 21%. This suggests that the quadratic model is

disproportionately magnifying the 6-bin data, implying the linear model is the least biased option; it also provides a more conservative effect estimate.

Table 3. Linear & quadratic model effects

Bins	Linear	Quadratic
Raw Data	1	1
4-Bin	0.96*	n/a
6-Bin	0.95	1.21

* Result not significant

Restating the process: we start with raw data, if we remove outliers (bandwidth z = ±1), then we have data that behaves. If we add bins, to remove noise, and use a linear model, the estimates remain stable. If we add a quadratic model to the bins, then we see distortion. The quadratic model appears too aggressive, removing noise but probably adding bias. The linear model is more likely conservatively smoothing the data. In addition, the 6-bin approach provides more control over the data. The graphing software outputs similar graphs for both raw and 6-bin sets of data, as the graph plotting algorithm will implement some form of internal bin-making for each of the repeated six rating (x) points. But what happens is unseen, on autopilot, within the application. If we provide the bins pre-made, then we have greater understanding and control over the plot outcome. The R-squares confirm this perspective. The raw data at bandwidth z = ±1 has an R-square of .09 (a very poor fit), while the 6-bin R-square at z = ±1 is .81 (a much better fit). For reference 1.0 is a perfect fit.

Sensitivity tests (step 7). Multiple tests were undertaken, to look at the impact of student results stretched across almost seven standard deviations (z-scores). The tests revealed that the full sample treatment effect is severely negatively impacted by scores beyond bandwidth z = -1. There are a mix of reasons: sample size, bin size, number of bins. But most importantly there is no counterweight to the extreme outliers on the control side of the graph (they are not controlled). All these reasons are combining to give too much weight to the outliers. However, inside of bandwidth z = ±1 the results appear to settle. The use of no bins, and evenly spaced 4-bin and 6-bin sets all resulted in very similar linear findings. Suggesting the underlying data that went into the three different mean score bin calculations was evenly dispersed and not impacted by width of the bin.

At this stage the approach appeared relatively conservative, bias had been minimized and some control asserted over software usage. The linear model fitted the dataset at bandwidth z = ±1 and the raw estimate was conservatively smoothed in a downwards direction.

2.6 Follow up survey

A very short purposive survey was conducted targeting the treatment group. The anonymous survey was stratified and sent separately via Google Form to three known

sub-strata; did not improve (DNI), improved (I), and super improvers (SI). The objective of the survey was to see if the Kahoot users were disproportionately found among the super improvers. It also inquired as to the impact of email reminders.

The survey was written in Japanese and asked four simple questions with multiple choice answer options.

1. How often did you practice for the quizzes using Kahoot?
2. For each quiz exercise there were three mini-games. How many did you practice?
3. How often did you open the email reminder from [course leader] about the quiz exercise?
4. How did you feel about the prompt emails?

The answers to questions one and two were combined into total plays, and then compared to known anonymized data on the treatment group actual Kahoot usage.

3 Results

A nonparametric linear 6-bin model was selected for estimation. 150 observations were used relatively evenly across six bins. The posttest data, when plotted against the pretest rating data, clearly shows an improvement at the cut-off point. The treatment effect was calculated and supported by two methods. Method one, it appears from visual inspection of Figure 4, that the point estimation is a z-score of about 0.42. Method two, the linear y-intercept formulas were calculated for the treatment bin data points as $y = 0.32 + 1.235x$ and the control bin data points as $y = -0.112 + 0.655x$. At x=0, the cut-off point, the treatment effect is $+0.32 - -0.112 = 0.432$. Since the standard deviation for the posttest was 3.52 points, the estimated treatment effect is 1.52 points.

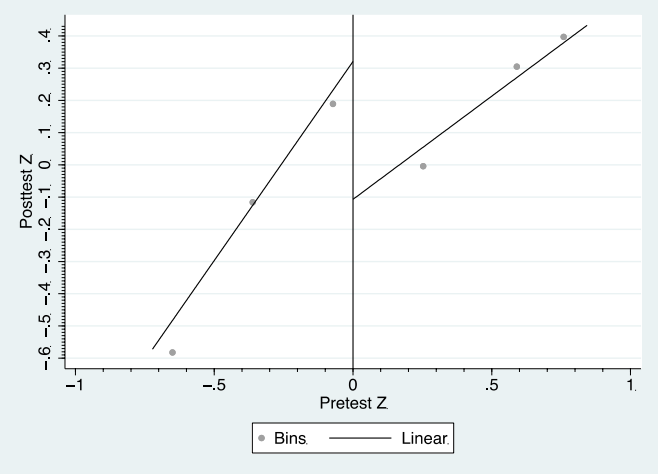

Fig. 4. Treatment effect

The RD treatment effect estimate is localized to the cut-off with a relatively small sample and school population. The estimate can only be tentatively extrapolated to other contexts. However, an average treatment gain of 1.52 points is unexpectedly high for the treatment, and it will be interesting to see if the result is replicated elsewhere. But why have the students gained? Is this the effect of gamification, nudging or a mix of both? The triangulation of RD posttest scores, Kahoot play data and treatment participant survey responses has provided answers.

3.1 Gamification play data

Evidently, the Kahoot usage data (Table 4) shows a lack of gamification activity, and this led to the formulation of a secondary hypothesis. Did the students who chose to use Kahoot improve more than those who did not? It is known that the Kahoot quizzes were played at most 105 times in combination by the 91 treatment group participants, and the first section of the first quiz, referred to as mini quiz 16-1, was played 35 of those times. Furthermore, Table 4 shows that no more than five students from 91 could have completed every quiz, and no more than seven students could have partially attempted all three quizzes. This is clear because mini-quiz 19-1 only has five plays and the quiz 18 mini-series, is the least used quiz series with a maximum of seven plays.

Table 4. Kahoot usage

Quiz	Usage	Quiz	Usage	Quiz	Usage
16-1	35	18-1	7	19-1	5
16-2	11	18-2	6	19-2	6
16-3	12	18-3	7	19-3	16

The Kahoot play was anonymous and possibly a small minority of students played all the games. The only way to know was to ask. To this end, a survey was sent out. The survey was anonymous, but three sub-strata were created based on known posttest gains. Since it was known that at the $z = \pm 1$ bandwidth, the 6-bin control group had regressed by $z = -0.298$, approximately one point, it seemed likely that the first point of the treatment group improvement was also positive regression to the mean. Consequently, the improvement strata were positioned from two points of gain. The super improvers were positioned at four points of gain.

3.2 The survey response

29 students (31.9%) from the treatment cohort of 91 responded to the survey. Interestingly, the response to the survey mirrored how well they had done on the quizzes. Only four from a potential 26 (15.4%) were from the did not improve (DNI) group, ten from 29 (34.5%) of the improved (I) group responded and 15 from 36 (41.7%) of the super improvers (SI) responded. It seems there is a strong correlation between the type of person who was willing to undertake a nudged action (answer a survey) and their actual quiz test improvement that itself was based on a nudged treatment. The survey

responses were equally illuminating and can be seen in Table 5. The 15 super improvers (SI) account for 34 of the 105 Kahoot plays (32.4%). The ten improved (I) respondents account for 7.6% of the plays and the four did not improve (DNI) respondents account for 1.9% of the plays. Only two of the 29 respondents claimed to have completed all nine mini quizzes, both were in the super improver (SI) grouping, and in isolation account for at least 18 plays. In fact, only 12 survey respondents (41.4% surveyed) tried Kahoot, and only eight surveyed respondents (27.6%) returned for a second week. On the other hand, the digital nudging results do correlate to the level of improvement. Most of the super improvers were opening the prompt emails. Indeed, six claim to have opened them every time. The improved group also tended to open the emails, but at lower rates. The did not improve grouping appear less likely to open the emails and this finding has been supported by the small number of DNI respondents to this survey.

Table 5. Survey data

1. How often did you practice for the quizzes using Kahoot?					
	All 3 Weeks	***2 Weeks***	***1 Week***	***Never***	
SI	2	4	2	7	
I	0	1	2	7	
DNI	0	1	0	3	
2. For each quiz exercise there were three mini-games. How many did you practice?					
	Three	***Two***	***One***	***Never***	
SI	4	0	4	7	
I	1	1	1	7	
DNI	0	0	1	3	
Combined Game Play Data					
SI		***I***	***DNI***		
34 plays		8 plays	2 plays		
3. How often did you open the email reminder from [course leader] about the quiz exercise?					
	Always	***Mostly***	***Sometimes***	***Rarely***	***Never***
SI	6	1	6	1	1
I	1	3	2	2	2
DNI	0	0	3	1	0
4. What did you feel about the prompt emails?					
	Strongly Positive	***Somewhat Positive***	***No Effect***	***Somewhat Negative***	***Strongly Negative***
SI	4	3	7	0	0
I	0	1	9	0	0
DNI	0	1	3	0	0

* Q4 SI one respondent did not answer.

4 Discussion

The Kahoot data suggests a need to unpack the hypothesis into two components: nudging and gamification. Gamification requires users to voluntarily participate. In this study, evidently most treatment group members decided not to participate. 2.2% of the students account for 17.1% of the plays. The two users (2.2%) achieved this by simply trying all the mini quizzes one time. This suggests that the other 97.8% of students' usage of the gamified materials was underwhelming, and in the majority of cases insignificant. Yet still, a 5% treatment effect was seen, implying the effect was mostly in the nudge.

4.1 The nudge as teaching strategy

The findings support a nudge teaching strategy. Teachers have some influence over their students and can nudge them towards new heights. This has always been true, but with digital choice environments, it is possible that nudging will be more evenly applied, as the platform environments will likely have in-built nudge mechanisms, even if the teacher does not proactively choose to use a nudge teaching strategy.

The findings appear to show that nudging users with official emails and a call to action was enough to see an improvement effect of 1.52 points or about 5%. For clarification, only six email messages from the course leader led to a 5% net average gain after accounting for gain accumulated through positive regression to the mean. This suggests strategic teacher action has influence on student outcomes. Furthermore, the results seem to suggest that students who were more receptive to being nudged into an action, such as trying gamified study or answering a survey, were more likely to take heed of the nudge and improve. It is not clear how they improved, only that nudging seems to have stimulated learning activity or greater focus, that has led to improvement.

4.2 Importance of voluntary participation

The findings support the gamification approach stated in [10], [13], as previously stated in the introduction. Essentially, active gamification cannot itself be nudged if the student does not want to do it. The users can be nudged to try. 41.4% of surveyed respondents were nudged to at least try, but they cannot be nudged to actively participate, as stated only 27.6% of surveyed respondents even returned for a second week of play. In the case of this research, the negative finding for gamification needs to be balanced against the fact that the students were the lower performing test takers, meaning they most likely included many students with less motivation to improve their score and/or English ability. Therefore, the data does not infer gamification does not work, it infers that gamification must be used with students who want to use it. It is a learning strategy more than a teaching strategy.

The key driver for magnified success is likely the individual student's decision to participate. If a student wants to participate, then the outcome is likely to be far better than if they solely do as their teacher asks. The implications of this, are that teachers must find ways to create student centred proactive learning experiences. Students must

be encouraged (nudged) to learn based on their own personal needs and desires. Once the students want to learn, then gamification as a learning strategy comes literally in to play.

The combination of nudging and gamification could have a powerful effect on learning outcomes for those who want to participate. The survey results showed the best performers, were often, though not exclusively, those who voluntarily elected to use Kahoot. The super improvers accounted for the bulk of the game plays. 29 surveyed players accounted for 44 game plays. 15 super improvers (51.7% of surveyed users) with gains of 4-12 points, accounted for 34 (77.2%) of those game plays. This might suggest learner desire to participate is key to significant improvement.

4.3 Limitations

The limitations of the study were primarily caused by COVID-19 distance education modifications to both teaching and learning. For example, the quizzes were online open book and with no real time limits. Students knew they could take their time. This limited their need to revise. They were de-incentivized to participate in extra study, even if it was gamified. Future research will demonstrate if this was a factor in the results. But since student choice is central to both nudging and gamification, any form of disincentive could not have been helpful. A second limitation was home study. If the test had been taken in the classroom, under test conditions, there would have been a lower mean and wider range of scores. This would have made RD implementation easier.

5 Conclusion

In this study an emphasis was placed on a libertarian paternalistic intervention. A technologically enhanced teaching strategy to nudge a gamified language learning activity towards those who had weaker results. The nudge seemed to have positive intervention effects, but the gamification did not. Further research is now required with an emphasis on offering help to L2 students who are known to want to study a second language. The hypothesis being that if you nudge a student to do something they want to do, they will improve. If you gamify that desired activity, the learning outcome will be further amplified. At this stage, further research is required to demonstrate in L2 contexts, but a combination of digital nudging and gamification could hold real promise for elective language learning study.

6 Acknowledgment

We would like to thank Fujishiro, Jinbou and Masui sensei for their kind co-operation.

7 References

[1] M. Jakubowski, 'Gamification in business and education: Project of gamified course for university students.', in *Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL Conference*, 2014, vol. 41, pp. 339–342. [Online]. Available: https://journals.tdl.org/absel/index.php/absel/article/view/2137

[2] 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.

[3] Y. Lam, K. Hew, and T. K. F. Chiu, 'Improving argumentative writing: Effects of a blended learning approach and gamification', *Lang. Learn. Technol.*, vol. 22, Feb. 2018.

[4] E. Zarzycka-Piskorz, 'Kahoot it or not? Can games be motivating in learning grammar?', *Teach. Engl. Technol.*, vol. 16, no. 3, pp. 17–36, 2016.

[5] H. T. T. Phuong, 'Gamified Learning: Are Vietnamese EFL Learners Ready Yet?', *Int. J. Emerg. Technol. Learn. IJET*, vol. 15, no. 24, p. 242, Dec. 2020. https://doi.org/10.3991/ijet.v15i24.16667

[6] K. Palaniappan and N. Md Noor, 'Gamification Strategy to Support Self-Directed Learning in an Online Learning Environment', *Int. J. Emerg. Technol. Learn. IJET*, vol. 17, no. 03, pp. 104–116, Feb. 2022. https://doi.org/10.3991/ijet.v17i03.27489

[7] S. Loewen *et al.*, 'Mobile-assisted language learning: A Duolingo case study', *ReCALL*, vol. 31, no. 3, pp. 293–311, Sep. 2019. https://doi.org/10.1017/S0958344019000065

[8] H. Bicen and S. Kocakoyun, 'Perceptions of Students for Gamification Approach: Kahoot as a Case Study', *Int. J. Emerg. Technol. Learn. IJET*, vol. 13, no. 02, p. 72, Feb. 2018. https://doi.org/10.3991/ijet.v13i02.7467

[9] 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.

[10] 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.

[11] 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

[12] J. Byrne, 'Anytime Autonomous English MALL App Engagement', *Int. J. Emerg. Technol. Learn. IJET*, vol. 14, no. 18, p. 145, Sep. 2019. https://doi.org/10.3991/ijet.v14i18.10763

[13] B. Burke, *Gamify: how gamification motivates people to do extraordinary things*. Brookline, MA: Bibliomotion, books + media, 2014.

[14] R. H. Thaler and C. R. Sunstein, *Nudge: improving decisions about health, wealth, and happiness*, Rev. and Expanded ed. New York: Penguin Books, 2009.

[15] M. Weinmann, C. Schneider, and J. vom Brocke, 'Digital Nudging', *Bus. Inf. Syst. Eng.*, vol. 58, no. 6, pp. 433–436, Dec. 2016. https://doi.org/10.1007/s12599-016-0453-1

[16] J. Byrne and M. Furuyabu, 'The Affordances and Troubleshooting of an IT Enabled EFL Classroom: Four Practical Examples', *Teach. Engl. Technol.*, vol. 19, no. 2, pp. 70–87, 2019.

[17] T. Okoshi, K. Tsubouchi, and H. Tokuda, 'Real-World Product Deployment of Adaptive Push Notification Scheduling on Smartphones', in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA, Jul. 2019, pp. 2792–2800. https://doi.org/10.1145/3292500.3330732

[18] D. N. Wagner, 'On the emergence and design of AI nudging: The gentle big brother?', *ROBONOMICS J. Autom. Econ.*, vol. 2, pp. 18–20, 2021.

[19] M. T. Damgaard and H. S. Nielsen, 'Nudging in education', *Econ. Educ. Rev.*, vol. 64, pp. 313–342, Jun. 2018. https://doi.org/10.1016/j.econedurev.2018.03.008

[20] A. T. Schmidt and B. Engelen, 'The ethics of nudging: An overview', *Philos. Compass*, vol. 15, no. 4, Apr. 2020. https://doi.org/10.1111/phc3.12658
[21] R. H. Thaler and C. R. Sunstein, 'Libertarian Paternalism', *Am. Econ. Rev.*, vol. 93, no. 2, pp. 175–179, Apr. 2003. https://doi.org/10.1257/000282803321947001
[22] P. Schochet *et al.*, 'Standards for regression discontinuity designs.', What Works Clearinghouse, 2010. [Online]. Available: https://files.eric.ed.gov/fulltext/ED510742.pdf
[23] A. Dicks and B. Lancee, 'Double Disadvantage in School? Children of Immigrants and the Relative Age Effect: A Regression Discontinuity Design Based on the Month of Birth', *Eur. Sociol. Rev.*, vol. 34, no. 3, pp. 319–333, Jun. 2018. https://doi.org/10.1093/esr/jcy014
[24] H. Dyson, J. Solity, W. Best, and C. Hulme, 'Effectiveness of a small-group vocabulary intervention programme: evidence from a regression discontinuity design: Effectiveness of a vocabulary programme', *Int. J. Lang. Commun. Disord.*, vol. 53, no. 5, pp. 947–958, Sep. 2018. https://doi.org/10.1111/1460-6984.12404
[25] A. Vardardottir, 'Peer effects and academic achievement: a regression discontinuity approach', *Econ. Educ. Rev.*, vol. 36, pp. 108–121, Oct. 2013. https://doi.org/10.1016/j.econedurev.2013.06.011
[26] W. C. Smith, 'Estimating unbiased treatment effects in education using a regression discontinuity design', *Pract. Assess. Res. Eval.*, vol. 19, no. 9, 2014. https://doi.org/10.7275/7911-vd52
[27] R. Jacob, P. Zhu, M. A. Somers, and H. Bloom, *A practical guide to regression discontinuity.* New York, NY: MDRC, 2012. [Online]. Available: http://www.mdrc.org/sites/default/files/regression_discontinuity_full.pdf
[28] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2001.
[29] A. Linden, J. L. Adams, and N. Roberts, 'Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design', *J. Eval. Clin. Pract.*, vol. 12, no. 2, pp. 124–131, Apr. 2006. https://doi.org/10.1111/j.1365-2753.2005.00573.x
[30] 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).
[31] J. A. Kaufhold, *Basic statistics for educational research: second edition.* Place of publication not identified: iUniverse Com, 2013.
[32] H. S. Bloom, 'Modern Regression Discontinuity Analysis', *J. Res. Educ. Eff.*, vol. 5, no. 1, pp. 43–82, Jan. 2012. https://doi.org/10.1080/19345747.2011.578707
[33] H. Luyten, 'An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95', *Oxf. Rev. Educ.*, vol. 32, no. 3, pp. 397–429, Jul. 2006. https://doi.org/10.1080/03054980600776589
[34] M. D. Cattaneo, M. Jansson, and X. Ma, 'Manipulation Testing Based on Density Discontinuity', *Stata J. Promot. Commun. Stat. Stata*, vol. 18, no. 1, pp. 234–261, Mar. 2018. https://doi.org/10.1177/1536867X1801800115
[35] StataCorp, *Stata Statistical Software: Release 17*. College Station, TX, 2021.
[36] M. D. Cattaneo, N. Idrobo, and R. Titiunik, *A Practical Introduction to Regression Discontinuity Designs: Foundations*, 1st ed. Cambridge University Press, 2019. https://doi.org/10.1017/9781108684606
[37] S. Calonico, M. D. Cattaneo, and R. Titiunik, 'Robust Data-Driven Inference in the Regression-Discontinuity Design', *Stata J. Promot. Commun. Stat. Stata*, vol. 14, no. 4, pp. 909–946, Dec. 2014. https://doi.org/10.1177/1536867X1401400413

8 Authors

Jason Byrne is an Associate Professor at INIAD, Toyo University, Tokyo 115-8650, Japan. Byrne has co-authored multiple 1 million download English study apps. His interests include CALL, digital nudging and gamification (email: byrne@toyo.jp).

Takehiko Ito is an Assistant Professor at Toyo University, Faculty of Information Networking for Innovation and Design (INIAD). He received a master's degree in Education from Keio University. His research interests include English communication and motivation (email: ito9041@toyo.jp).

Mariko Furuyabu is an Associate Professor at Toyo University, Faculty of Information Networking for Innovation and Design (INIAD). She received her doctorate in Linguistics from Sophia University. Her research interests include language learning with technology, and second language pragmatics (email: furuyabu@toyo.jp).

Article submitted 2022-03-02. Resubmitted 2022-03-29. Final acceptance 2022-03-29. Final version published as submitted by the authors.