Encouraging Errors Through Gradual Feedback to Improve Vocabulary Learning

Lukas Ansteeg[a], Ton Dijkstra[b] and Frank Leoné[c]
Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands

Keywords: Error Commission, Feedback, Orthographic Similarity, Representational Similarity, Word Learning.

Abstract: Making errors that are related to the correct answer can aid the learning of vocabulary. Encouraging error commission may improve learning outcomes and provide insight into latent learning processes. We investigate the possibility of eliciting useful errors through the use of orthographic similarity-based feedback. We find that error commissions fully replace non-answers during a learning task and largely replace them during post-tests. Participants receiving similarity-based feedback seem to better consolidate orthographic knowledge over a one-week delay. The committed errors provide evidence for gradual learning of sublexical elements and for theories holding that specificity of representations increases during learning. Gradual feedback shows to also have motivational benefits. These findings suggest promising insights for classroom and digital vocabulary instruction.

1 INTRODUCTION

When learning a new word in a second language, we usually need multiple attempts and make errors along the way before fully knowing the new word. Can those errors give us more insight into how new words are learned? And if so, can this insight help us organize learning in a better way? Evidence suggests that under some circumstances producing errors can help with learning a new word (Metcalfe, 2017). This requires feedback from a teacher or learning software that helps us to correct errors and stimulate learning. In this study, we will investigate whether feedback can encourage learners to commit more errors when they would otherwise give no response at all, and whether this in turn improves performance. We also consider if learning occurs spontaneously or gradually, and whether error commission can improve learner models for intelligent tutoring systems. Setting the stage for our experiment, we will first look at how words are internally represented, then at how errors made by learners can expose these latent representations and, finally, how feedback is used to correct these errors and strengthen the correct representations.

1.1 Errors and Internal Representations

How words are learned is an ongoing subject of debate. We do know that knowledge of vocabulary is not binary; a word is not simply unknown or known to the learner (Laufer & Goldstein, 2004; Palmberg, 1987). Several theories have suggested, for example, that knowledge of a new word has to be consolidated to be remembered in the long term (Bakker et al., 2014), that the likelihood of remembering a learned word can be modeled as a function of intervals between learning episodes (Settles & Meeder, 2016), and that words might be known receptively but not freely produced by the learner (Laufer, 1998; Laufer & Paribakht, 1998). These are not competing accounts of how learning occurs but largely complementary theories, the number of which illustrates the complexity of the learning process.

Lexical items, or words, in turn are not indivisible units. In fact, it is debatable where to draw the line of what constitutes ‘one unit’ (Bogaards, 2001; Brown et al., 2022). For simplicity, we will define a word as a form-meaning unit consisting of a single orthographic form and a linked semantic part for the

[a] https://orcid.org/0000-0003-1202-6235
[b] https://orcid.org/0000-0003-2514-5866
[c] https://orcid.org/0000-0002-3703-6504

Ansteeg, L., Dijkstra, T. and Leoné, F.
Encouraging Errors Through Gradual Feedback to Improve Vocabulary Learning.
DOI: 10.5220/0011975100003470
In *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023) - Volume 1*, pages 79-91
ISBN: 978-989-758-641-5; ISSN: 2184-5026

purposes of this study, disregarding phonology, word families, and morphological properties. In terms of orthography, words are further made up of sublexical elements such as letters. In order to learn a new word, the learner must internalize and connect all these aspects. What they know, or think they know, of a word is considered their mental representation. It is sometimes also referred to as the latent representation, since we can usually only indirectly assess the internal state of a student by their behavior. A student learning Finnish who when asked to complete the word 'taiv_s' correctly responds with 'taivas', but is unable to explain the meaning of the word or provide a translation equivalent, cannot be said to have fully learned the word. Nevertheless, they clearly have knowledge of the orthographic form and thus partial knowledge of the word.

As a gradual measure of how well a learner knows the orthographic form of a new word, one could measure how similar the correct word form and their best guess for the word are. If two students are learning that the Finnish word form for 'air' is 'taivas', the student guessing 'taivos' can be said to have better learned the orthographic form than a student guessing 'taippu', even though both are incorrect. In this case, their representations of the word can be said to be underspecified (Janssen et al., 2015). A useful tool to express this is in terms of Levenshtein distance (Levenshtein & Others, 1966), a metric assessing the similarity of two-word strings (e.g. Schepens et al., 2012).

Several models of lexical usage and learning account for the multiple modalities of words (semantic, phonological, orthographic) (i.e., Dijkstra & van Heuven, 2002); an attempt to explain the interplay between modalities and underspecified representations during learning is made by the ontogenesis model (Bordag et al., 2022). These theories suggest that learning a word is the process of internalizing all the constituent parts that make up the correct and complete representation of the word, and that this process can be gradual rather than all at once. This has important implications for education, because a gradual, stepwise process of word learning may imply that learners can partially know words and only need to focus on other parts of the words to complete their learning.

Although psycholinguistic models show mental lexical representations to be complex, and educational research assumes word learning to be incremental and gradual (de Groot, 1995; Groot, 2000), little research has been done about the degree to which second language lexical representations are formed spontaneously or gradually in the sense of their constituent parts. We will be referring to gradual learning as piecemeal, part-by-part learning of representations in which not just memory is strengthened or consolidated over time, but the representation itself becomes more clearly defined. In this gradual process, representations are sharpened from vague and overlapping to clearly defined and distinct (Baxter et al., 2021; Davis & Gaskell, 2009; De Grauwe et al., 2014), also referred to as lexical specificity (Janssen et al. 2015).

Knowing how well learners know any given word is important for tracking their overall progress. This allows teachers or digital learning software to give the learners appropriate feedback, make better decisions about further learning activities, and to motivate learners by making their progress visible (Hooshyar et al. 2020). However, the latent representations of newly learned words within a learner's mental lexicon are by definition hidden from us. We can only measure their knowledge via responses to learning tasks and tests. When learners are not encouraged to respond in tasks where they do not know the answer, information about their knowledge is binary; they either know or do not know. When learners are allowed or even encouraged to make errors, on the other hand, we can gather information about why they may have made a particular error and thus estimate flaws in their developing internal representations. Encouraging error commission may thus increase the ability to model learner knowledge by revealing information about partially learned words, for example, by revealing whether some letters or syllables of the item are already known. This is one of the hypotheses we intend to test in this paper, but in order to do so, we must find a way to encourage learners to commit more errors. As our ultimate goal is to improve language learning, we will first look at the role of errors on learning, since inducing errors in participants at the detriment to their learning would defeat the purpose.

1.2 Errors and Learning

Learning tasks for new words can be errorless or errorful. Listening to a native speaker or reading vocabulary from a word list is inherently errorless, as learners are exposed only to the correct L2. In contrast, a learner may make errors when producing new words, for instance during retrieval practice. When deciding whether allowing learners to produce errors is conducive to learning, there are two possible approaches. First, errors could be minimized, because only exposure to or production of correct responses will strengthen the representation of the correct word.

Second, errors could be encouraged, because they allow learners to reflect upon and correct their assumptions, thus paying additional attention and cognitive effort to the word. Both approaches have evidence in support of them. Some situations favor errorless learning, given that error production is likely to consolidate the erroneous response (Wilson et al., 1994). In others, error generation has been shown to be beneficial to learning if errors are followed by corrective feedback (Kornell et al., 2009, 2015). One example of effective errorful learning is retrieval practice, which is the attempt of reproduction of something learned and thereby strengthening existing representations (Karpicke, 2017). Evidence from Potts et al. (2019) suggests a beneficial effect of 'guessing' on vocabulary retrieval.

A crucial requirement of useful errors seems to be that a committed error and the correct response need to be related; random guessing does not lead to better learning outcomes (Metcalfe, 2017). In this paper, we will therefore refer to incorrect responses as error commission rather than guesses, so when talking about errorful learning or encouraging errors, we refer to encouraging participants to express their erroneous assumption instead of not responding at all.

The literature on learning from errors is largely based on semantically related words. However, the same benefit of error commission might be present when errors are orthographically related. Similarity of committed errors to the correct response may be useful for the same reasons that contrasting orthographically similar words is beneficial for word learning (Baxter et al., 2021). Not only do the error and correct response share underlying sublexical features and thus may have overlapping representations, but the corrective feedback may highlight differences between the current representation and the correct one, allowing the learner to adjust their assumptions. Orthographic neighbors, that is words that are spelled similarly, are often confused by learners, but can also build on each other and thus be learned more easily when the learner consciously distinguishes them (van Heuven et al., 1998). While erroneous responses during word learning can be random misspellings of the correct response, they can also be based on confusion with other similarly spelled words. Form-focused retrieval practice should thus generate orthographically related errors.

Retrieval practice with corrective feedback thus satisfies our requirements: It generates learner errors and is an effective learning strategy. The literature suggests that related errors might increase learning outcomes, so we will test our second hypothesis: that an increase in error commission results in better learning outcomes. Maximizing error commission may thus benefit the learner both in the short term by increasing learning gains and in the long term by allowing teachers or learning software to better adjust future learning tasks. This only leaves a mechanism for increasing error commission. Given that corrective feedback is inherently needed for retrieval practice, we want to investigate whether the type of feedback given can encourage the production of useful errors.

1.3 Errors and Feedback

We are interested in whether the form of corrective feedback to learner errors can be manipulated to increase error commission.

The first step to encouraging error commission would be to make them less discouraging. Creating an environment where making errors is not penalized formally or socially is crucial for language teachers (Young, 1991). While social inhibition might play less of a role in learning software, users are still influenced by the affective nature of the feedback they receive (Moridis & Economides, 2008). Making errors can inherently be frustrating even when they are not penalized (Heimbeck et al., 2003; Tulis & Ainley, 2011). A positive response to incorrect answers, praising either the attempt or the parts of the answer that were correct even if the whole was not, can result in an overall more positive experience for the learner (Talmi, 2013). Thus, corrective feedback to errors should be positive and focus on the learning value of the error or the achievement of the attempt, rather than on the negative aspects.

Extrinsic rewards are simple and effective in guiding user behavior in learning software (Filsecker & Hickey, 2014; Gooch et al., 2016). In a scored learning task, simply rewarding points for any given response is likely to encourage users to commit a response regardless of whether they know the answer or not. Our goal is to increase error commission without loss of learning gain. A partial score for incorrect answers and full score for correct answers should lead to participants learning words while discouraging non-responses when the learner is unsure about the answer. A study by Abraham et al. (2019) has shown a positive impact of rewards during retrieval practice with semantically related prompts.

Because errorful learning is most conducive to learning when errors are related to the learning target, we can go even further and specifically encourage 'useful' errors. If the learning of underlying representations is gradual, given responses during learning should gradually approach the correct

answer. As we expect learning gains from errors based on their orthographic similarity to the correct answer, we can reward participants with points proportional to how similar their error was to the correct answer.

Highlighting a partially correct answer as partially correct may also help the learner pinpoint where they went wrong, rather than invalidating their entire response. Even on a single word level, such gradual feedback could account for a range of 'correctness' rather than a binary correct / incorrect response. This leads us to our final hypothesis, that gradual feedback can increase error commission.

1.4 Current Experiment

In the present study, we test the effect of gradual feedback on learner behavior and task performance. We conducted a word learning experiment with Finnish as the target language as it conforms to the same script as known by our native Dutch participants, but shares no common origin with Dutch words. We thus avoid the effects of cross-linguistic similarity.

We will test three hypotheses. First, we expect gradual feedback to result in more error commission (as opposed to non-submission) than binary feedback. We test this by explicitly showing the participants how close they are orthographically to the correct answer by measure of Levenshtein distance and rewarding them partial points for partially correct responses. We expect this to encourage error commission in situations where learners would otherwise commit no response. We compare this gradual feedback condition to a binary feedback condition in which participants are only informed whether their response is correct or incorrect.

Second, we anticipate that participants in the gradual feedback condition will score higher on a post-test and delayed post-test. Assuming that the erroneous responses elicited by gradual feedback will be similar in orthography to the correct response, we expect this to increase the number of learned words as measured by the post-tests.

Third, we predict that words are learned gradually rather than abruptly, accompanied by a gradually decreasing Levenshtein distance of given answers to the correct answer over learning blocks. We expect the learning of underlying representations to be gradual in both conditions, but anticipate to find more evidence of gradual learning in the gradual feedback condition through increased error commission.

2 METHOD

2.1 Participants

Eighty-one participants completed two online word learning sessions. All participants were native speakers of Dutch recruited through the Radboud University SONA participant system and compensated with study credits or 17.50 euros for participation. Participants gave informed consent in accordance with guidelines of the Radboud University Social Sciences Ethics Committee (ECSW-2018-115).

The data of one participant were excluded from further analysis for failing to learn a single word. Therefore, in total 80 native Dutch speakers (67 female, mean age: 22.1 years) were included in the analysis. There were 41 participants in the binary condition and 39 in the gradual condition.

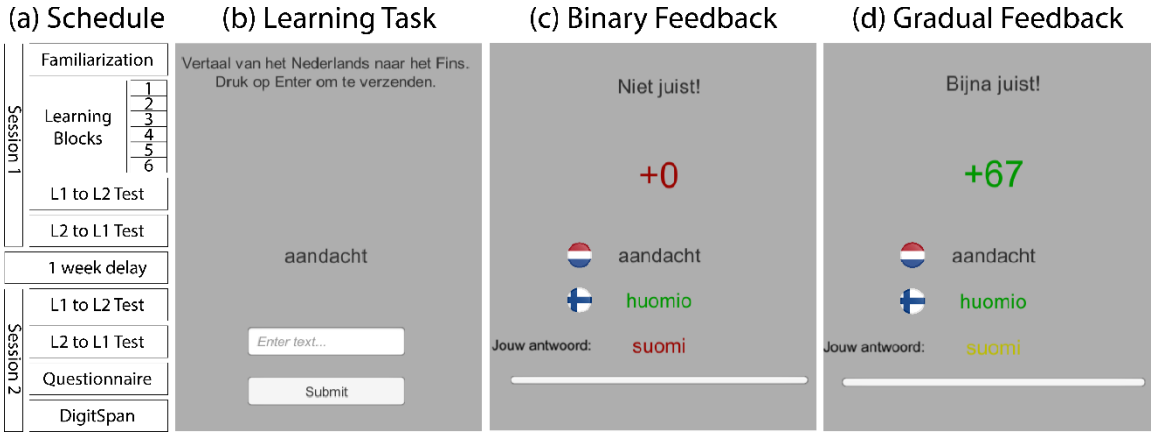

Figure 1: (a) Experimental design and schedule of the two sessions. (b) An example learning block task (b) and the feedback a participant receives in the (c) binary or (d) gradual condition. For (b) to (d) only the center of the screen is shown; the screen also includes a point total and instructions button.

2.2 Materials

Participants took part in the online experiments on laptops or PCs with a screen of at least 13" diagonally, using trackpad or mouse to interact with the experiment. Thirty Finnish proper nouns and adjectives were selected for which both the Finnish word and the Dutch translation equivalent were between 5 and 8 letters long. Half of the Finnish words have one or multiple orthographic neighbors, as defined by a Levenshtein distance of up to 2, within the set. Words including letters not found in Dutch were excluded. For the word list and more details on item selection see the appendix.

2.3 Procedure

The experiment consisted of an online learning session and a one-week delayed post-test. Participants were assigned to one of two presentation conditions: Binary feedback or gradual feedback, resulting in a mixed 2x2 design. In the learning session, participants were first familiarized with 30 Finnish words by seeing them together with their Dutch translation for 10 seconds per word. They then completed 6 blocks of word learning through a L1 to L2 translation typing task followed by corrective feedback. Participants had unlimited time for translations, and saw the feedback for 8 seconds. Each block contained every word once in an order pseudo-randomized for each participant. When given corrective feedback, participants received points for their answer. In both conditions, 100 points were given for a correct response. In the binary condition, 0 points were rewarded for incorrect responses. In the gradual condition, participants could receive points even for incorrect responses, depending on how close their response was to the correct response in terms of normalized Levenshtein distance, where P are the received points, lev is Levenshtein distance, len is word length, A is the given answer and T the target answer.

$$P = 100 - 100 * \frac{lev(A,T)}{max(len(A),len(T))} \quad (1)$$

Points were added up throughout each block and shown at the end of each block in comparison to the previous block's score to highlight improvement. The learning session ended with an L1 to L2 recall test in the format of the learning task, but without corrective feedback, and an L2 to L1 recognition typing task. While the L1 to L2 offline translation task (recall) tests learners' productive knowledge of the word, the L2 to L1 offline translation task (recognition) tests the receptive knowledge of the word. The order is chosen to minimize testing effects, under the assumption that most words that are known productively would be known receptively in any case (Laufer & Paribakht, 1998).

The delayed post-test took place one week (between 144 and 192 hours after completion of the learning session) and consisted of the same recall and recognition tests. It was followed by a short questionnaire collecting age, gender, educational status, language background, and the participant's perceived enjoyment of the learning task on a scale from 1 to 7. Participants then completed a digit span task to assess verbal memory capacity according to the method described by Woods et al. (2011). The digit span task is used to account for inter-individual variability in word learning capacities. In the first round of this task, participants saw 14 series of numbers digit by digit, and were then asked to reproduce the number. On success the length of the series increases by one, whereas two failures reduce the length by one. In the second round, they are asked to reproduce the series in reverse.

3 DATA PREPROCESSING

We will first explain the variables used in the linear mixed models we apply to answer our hypotheses about error commissions and learning outcomes and then elaborate on how we coded for gradual learning.

Table 1: Word learning examples from experiment participants.

Example	Word	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Max Improvement
A	pohjola	huopios	vojoh	pohjovo	pohjola	pohjola	pohjola	0.286
B	kasvot	kasvot	kasvot	kasvot	kasvot	kasvot	kasvot	1
C	kaava	-	-	-	kaava	kaava	kaava	1
D	osuma	-	-	-	esema	osuma	osuma	0.6
E	varjo	-	-	-	volga	varjo	varjo	0.8

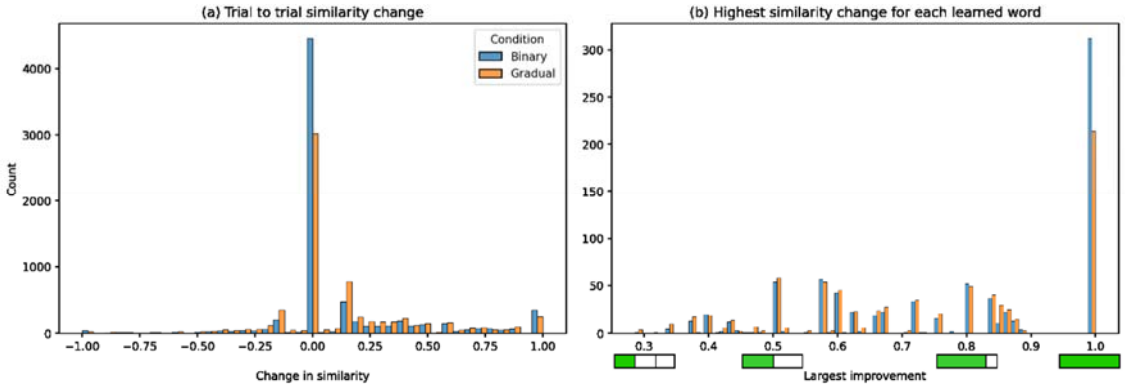

Figure 2: (a) The distribution of all trial-to-trial changes in similarity of given answers to correct answers from one learning block to the next. (b) The single highest trial-to-trial change for each word that was correctly answered in the final block by a participant. The green bars below the x-axis exemplify the interpretation of these values: For 0.33, at least two more equal or smaller learning steps must have occurred; for 0.5 at least one more; for 0.8 at least one additional small step; and for 1 the word was learned within a single trial. Note that for histograms (a) and (b) each pair of binary and gradual condition bars represent one bin range together. Though we split the data by condition, this graph is not meant to highlight a particular difference between the conditions, but rather to visualize the general distribution of results.

3.1 Error Commission and Learning Outcomes

We tested our hypotheses about error commission and learning outcomes using linear mixed models (LMMs) based on guidelines by Meteyard and Davies (2020). Analyses were run using the lme4 package in R (Bates et al., 2014) with the logit link (binomial) function given the binomial nature of the data on the trial level (correct or incorrect). We included in analyses all variables relevant to our hypotheses:

As outcome variables, we use the binary variable error commission to denote whether an incorrect answer was given (as opposed to none) to answer our first hypothesis. Models for our second hypothesis use the binary outcome result: correct or incorrect.

The fixed effects used include condition and digit span score. We also include time, which in the case of learning trials denotes the block from 1 to 6, and for post-test denotes immediate or delayed.

Separate models were run for learning outcomes (result) for the recall and recognition tests rather than including test type as a variable, as we consider them to make different task demands in regard to word knowledge.

3.2 Gradual Learning

Testing our hypothesis that word learning is gradual is less straightforward than our other two hypotheses. We hypothesize that latent word learning is always gradual, that each exposure to the new word and each attempted production strengthens and specifies the mental representation of this word. Of course, this latent learning process is not readily accessible to us in behavioral tasks (Preacher et al., 2008); all we see is the best guess a participant has at the time of a production task, if they are confident enough to attempt an answer in the first place. We have devised the following measure for word learning trajectories of participants based on their improvements from block to block in terms of Levenshtein distance to the correct answer.

For each word that a participant learned, we recorded the maximum improvement between any two blocks in Levenshtein distance, considering the word to have been learned gradually if this value is low and spontaneously if the value is high. Given that we wish to show that learning is gradual, we biased this measure toward spontaneous learning, so that no ambiguous cases could be falsely counted as implying gradual learning.

Table 1 shows example learning trajectories from our data. Example A is an ideal example of gradual learning. In the first block, the given answer is barely related to the correct answer. In each following block (2, 3, and 4), the answer improves slightly compared to the previous block, with block 4 showing the first correct answer. Because the three steps between the first four blocks each correspond to a 0.286 improvement in Levenshtein distance to the correct answer, we interpret this as three incremental learning steps.

On the other hand, we have B as a clear example of spontaneous learning: The participant answered correctly from the start, having learned the word during familiarization, and makes no mistakes thereafter. Example C shows a word that is abruptly

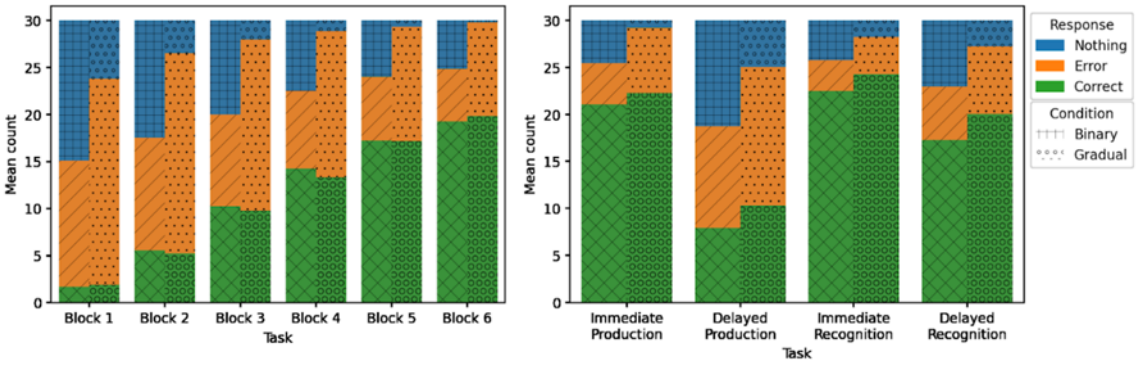

Figure 3: The mean correct, incorrect, and non-responses for each learning block on the left, and each post-test on the right. For each task, the left bar represents the binary condition and the right bar the gradual condition.

known after several non-commissions. It could be interpreted to have been gradually learned internally by a participant who was not confident to commit an answer until they were sure they knew it, or as spontaneously learned between block 3 and 4. Both cases are counted as spontaneous learning.

However, most observed cases are less clear cut, such as D and E, where after several non-commissions the participants makes an erroneous response and then the correct one. In our analysis we count D as gradually learned, with a step of 0.4 and 0.6, but E as spontaneously learned with steps of 0.2 and 0.8. Although both cases could be argued to involve at least two gradual learning steps, we do not count case E as gradual learning to exclude the chance that random guesses or mere orthotactic or statistical knowledge are responsible for the 0.2 improvement (cf. Storkel et al., 2006).

This table of examples leaves out many edge cases that occur during learning, such as words that were almost learned, where the final response is only a slight misspelling of the correct answer, or words that were learned, then briefly forgotten, then answered correctly again. Given that we want to test our hypothesis of gradual learning, we chose our analysis to attribute any ambiguous cases to count against our hypothesis. We thus chose to reduce the data by counting only the highest single improvement for each learned word for each participant, which entails that all other trial-by-trial changes for that word were equal or smaller (seen below shown in the change from Figure 2a to 2b). Counting all trial-to-trial changes would introduce the theoretical possibility that a word that was spontaneously learned in the last block would be counted as gradually learned due to small detected ‘improvements’ which are actually the result of random guessing in early blocks. As seen by the negative values in figure 2, participants also frequently show small regressions. While this observation is not theoretically opposed to learning being gradual, it could be interpreted as participants simply guessing randomly, with some answers being closer or further away from the correct answer by chance.

We selected a threshold of 0.75 as the maximum for gradual learning to be generous towards the hypothesis that spontaneous learning occurred. Even the largest trial-to-trial improvement under this threshold will be at least two letters off from the target word, requiring one or more additional steps until the word is fully correct. Cases like example E, where the participant first guessed ‘volga’ and then, correctly, ‘varjo’, are still counted as spontaneous, even though some evidence of a gradual step between can be seen.

4 RESULTS

The 41 binary condition participants and 39 gradual condition participants did not significantly differ in age, educational achievement, or verbal memory as shown by an insignificant t-test on the results of the digit span task.

The gradual condition was found to be more motivating overall. On a Likert scale from 1 to 7, participants in the gradual condition reported a mean enjoyment of 5.36 (SD = 1.10), compared to a mean enjoyment of 4.34 (SD = 1.24) in the binary condition ($t(78) = 3.83$, $p = .0003$). A free text prompt asking about the task mirrored this result, with most participants in the gradual condition reporting enjoying the gradual feedback, though some also reported finding it confusing.

4.1 Error Commission

As our first hypothesis, we proposed that gradual feedback leads to higher willingness to commit errors

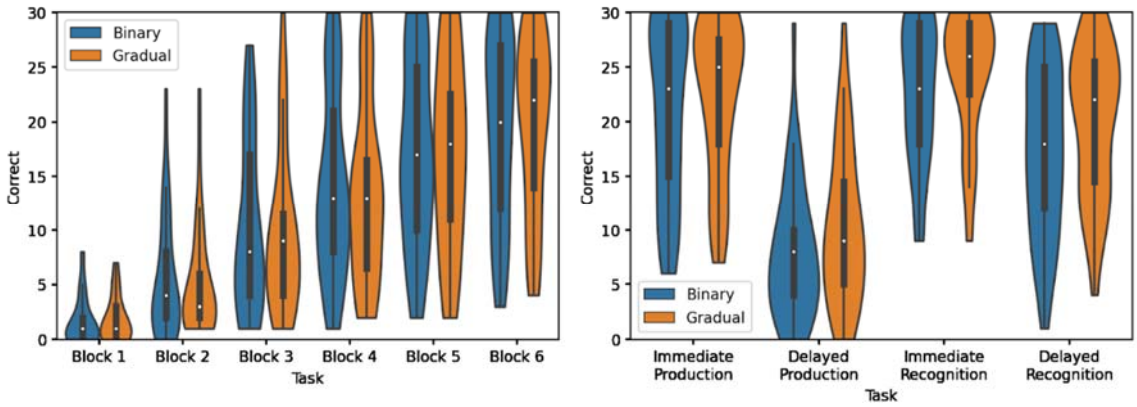

Figure 4: Violin plot of correct responses by condition during learning trials (left) and tests (right). The white dot shows the mean; the distribution of participant scores is shown by the violin shape.

in learning tasks, in contrast to skipping to the next item without committing an answer. Figure 3 leaves little doubt about whether the gradual condition encourages more error commission: In the binary condition, non-commissions and error-commissions are roughly even across all blocks; however, in the gradual condition subjects committed significantly more errors already in the first block, and non-commissions are practically non-existent by the last block. This difference carries over even into the post-tests, where error commission was no longer encouraged by instructions or point rewards.

We tested our hypothesis with three linear mixed models of error commission vs non-commission between conditions for all trials without a correct answer. For a complete output of the analysis, please see the appendix. All three models used the following formula:

$$\text{result} \sim \text{time} * \text{condition} + \text{digit span} + (1|\text{participant}) + (1|\text{word}) \quad (2)$$

We trained this model separately with data from the learning blocks, the recall test, and the recognition test. During learning blocks we see a strong positive effect of the gradual condition on error commission (beta=2.00, p<.001). There is a significant main effect of time (beta=.26, p<.001) and a positive interaction with condition (beta=0.25, p<0.001). This indicates that error commission increased over time for both conditions, but did so more sharply for gradual feedback participants.

The gradual condition also resulted in significantly more error commissions in the recall tests (beta=2.13, p<0.001), both immediate and one-week delayed. Error commission decreased as a main effect of time (beta=-.98, p<.001). For the recognition test, participants write down L1 words, testing their associative knowledge, which may not be affected by the form focused gradual feedback. Though we expected no difference in error commission here, we find a nearly significant effect of condition (beta=1.23, p=.064) and a barely significant interaction between time and condition (beta=.65, p=.047).

4.2 Learning and Test Performance

Our second hypothesis was that committing more errors would lead to an overall better learning outcome. As can be seen in figure 4, participants varied highly in their performance. During the first block of learning trials following familiarization, the median correct responses were 1 out of 30, although a few participants had already learned several words from familiarization (M = 1.81, SD = 2.09). By the sixth and final round of learning trials, participants had learned on average 19.52 words (SD = 8.18). Although the median seems to have diverged by block 6, there is no significant difference in performance between the binary condition (M = 19.24, SD = 8.43) and the gradual condition (M = 19.82, SD = 7.91).

For immediate and delayed post-tests, performance was similarly diverse as seen on the right side of figure 4. The mean performance is higher for the gradual feedback condition across all four tests, and particularly for the delayed tests. However, individual variance was high, so poorly and well performing subjects are found in either condition. We ran one linear mixed model each for the recall and recognition tests:

$$\text{result} \sim \text{time} * \text{condition} + \text{digit span} + (1|\text{participant}) + (1|\text{word}) \quad (3)$$

Performance in the digit span task accounted for much of the individual differences in word learning

performance for recall (beta=.40, p=.011) and recognition (beta=.39, p=.015). As expected for any learning task, performance dropped over time for recall (beta=-2.68, p<.001) and recognition (beta=-1.16, p<.001). Neither for recall nor for recognition was the condition a significant main effect. However, in the recall test a significant positive interaction effect with time arose (beta=0.34, p=0.037), implying that gradual condition participants did better in the one-week delayed test compared to the immediate post-test than binary condition participants.

4.3 Gradual Learning

Our third hypothesis is that words are learned gradually rather than abruptly. Although we split the plots in figure 5b by condition, we are also interested in the combined distribution, as our hypothesis states that word learning is gradual in general and not solely elicited by the gradual feedback condition. We consider the existence of words learned solely in steps of 0.75 and lower (left side of 5b) as evidence for gradual learning, while words above the threshold may be said to have been learned 'spontaneously' from one trial to another.

While the thresholded data answers whether gradual learning occurs in general, a comparison between conditions on the non-thresholded data seen in figure 2b reveals the effect of the manipulation on the presence of gradual learning steps. The average largest trial-to-trial improvement in the gradual condition ($M = 0.743$, $SD = 0.207$) was significantly lower than in the binary condition ($M = 0.788$, $SD = 0.207$): $t(1560) = -4.240$, $p < 0.001$). This could be interpreted as either learning being more gradual when gradual feedback is given, or, as we will argue in the discussion, as finding more evidence of underlying gradual learning progress through the increase in error-commissions.

5 DISCUSSION

We will discuss the results in the order of our three hypotheses: First, we set out to answer whether gradual feedback increases error commission. We found a strong effect, not just in the learning task itself, but even in one-week delayed tests where no incentives for error commission were given. Second, we asked whether gradual feedback results in better learning outcomes. We found an interaction with time but no significant main effect. Third, we sought to answer whether words are learned gradually and whether we can see this reflected in the error commissions. Our data provides evidence in support of this view. At the end of the paper, we will reflect on the relevance for both education practice and follow-up research.

5.1 Does Gradual Feedback Increase Error Commission?

As our first hypothesis, we predicted that gradual feedback would increase error commission. We did indeed obtain clear evidence in support of this view. The gradual feedback condition elicited strongly increased error commission from the first learning trial to the delayed post-test. Given that there was a large difference in error commissions already in the first learning block, even just mentioning point rewards for partially correct answers seems to affect task behavior. A habit of committing errors when unsure of the correct answer might already start to build up throughout the first block, as each trial allocates partial points for committed errors. Although participants in neither condition were penalized for failed attempts, the instructions including rewards for incorrect answers may have created an even clearer experience of a penalty-free environment and thus have led to higher engagement as described by Young (1991).

Throughout the learning blocks, the gradual feedback strengthens this habit to attempt to answer even if unsure. Error commission in the gradual condition increases steeply until participants virtually never omit their answer by the final block. This is in line with the observation of Abraham et al. (2019) for semantic similarity-based rewards and shows that rewarding based on orthographic similarity can produce the same incentive for error commission. Given that producing a partially correct word while learning vocabulary is a frequent occurrence, this creates a simple and effective mechanism for encouraging error commission.

Given that the post-test is very similar to the learning task aside from the lack of feedback, it is perhaps not too surprising that this habit to commit errors continues into the post-test, even though it is no longer explicitly encouraged and in no way rewarded. The observation that participants of the gradual condition still are much more likely to commit errors in the one-week-delayed post-test, however, suggests that at least a medium-term habit was formed during the short learning phase. It may create a long-term habit if learners internalize the intrinsic value of retrieval and error commission (Lally & Gardner, 2013). It would be interesting to investigate whether this effect would translate to

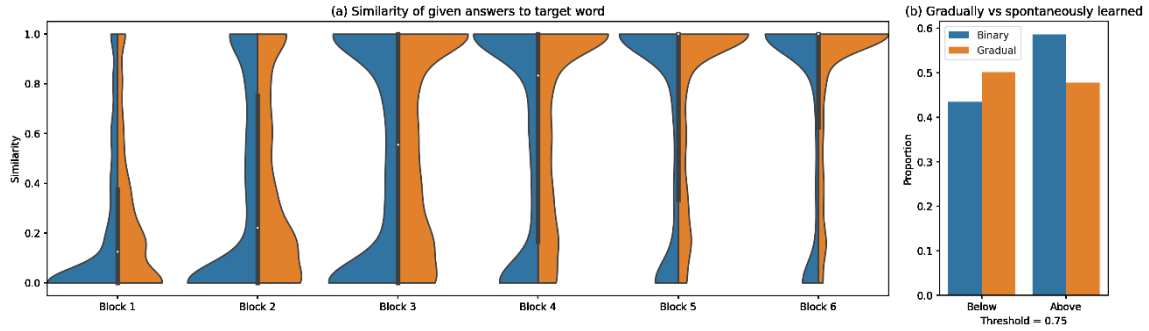

Figure 5: (a) The distribution of similarity (1 - Levenshtein distance) of given answers to correct answers, per learning block and split by condition. Instantaneous learning would be reflected by items only at the ends of the distribution, with completely unknown words at similarity 0 being replaced by fully known words at similarity 1 from block to block. Gradual learning is reflected by items 'trickling up' from unknown to partially known before settling at fully learned. (b) The distribution from figure 2 is split into two bins by the threshold of 0.75. Words below the threshold are considered learned gradually, and words above potentially learned spontaneously.

different learning tasks done by the same subject and whether a binary feedback task would 'reset' the learner to their original unwillingness to risk errors.

In the present study we only measured the responses themselves. Other studies have explicitly measured participants' confidence in their answers for each trial (e.g., Butler et al., 2011; Butterfield & Mangels, 2003). We omitted such a measure to retain the flow of the learning phase and restrict its duration, but a follow-up experiment including such a metric would shine more light into the decision making behind responses and non-responses. An interesting question, for example, is whether increased confidence to commit answers even when uncertain, might result in occasional correct responses where a less confident participant might omit their equally correct guess.

5.2 Does Increased Error Commission Result in Better Learning Outcomes?

In our second hypothesis, we predicted that error commission would intrinsically aid learning by encouraging active retrieval practice. We assumed that if the gradual feedback would elicit more error commission, it would also result in higher test scores. Numerically this does seem to be the case, with gradual feedback condition participants on average scoring two words better on the delayed production, and three words higher on the delayed recognition test. While those effects exceeded our expectations, unfortunately so did the extreme variance between participants, which leaves us with insufficient power to detect a significant effect. A replication in a more controlled environment than our online experiment, or a within-participant manipulation, may be able to confirm the seemingly high difference.

The significant interaction effect between the gradual condition and time reveals that gradual condition participants show a smaller decrease in learned words over the one-week period between sessions. This was only found in the form-focused recall test, but not in the recognition test. This may be the result of gradual feedback leading to different learning strategies or learning effects which are more consistently consolidated after time. This would be in line with findings by Baxter et al. (2021) that orthographic contrasting can lead to better long-term retention. It is also in line with the known phenomenon of competition between orthographic or phonological neighbors only occurring after consolidation (sleep) (Bakker et al., 2014), if we assume that the form-focused feedback of the gradual learning condition makes learners focus more explicitly on the orthographic form.

5.3 Are Words Learned Gradually?

The tests regarding our third hypothesis show that at least about half of all words were learned in multiple, gradual steps rather than at once. While latent learning may well be even more granular than observed (Daw & Courville, 2008), this leaves little doubt that word representations can be learned partially. Words were not just learned 'gradually' in the sense of certainty about the answer, but stepwise on a sublexical level. The data therefore adds evidence that vocabulary is not just stored in constituent parts which overlap between similar words, as suggested by models such as BIA+ (Dijkstra & van Heuven, 2002), but also that these parts are learned piecewise as suggested by the

ontogenesis model (Bordag et al., 2022). Our data strongly supports the theory that word learning relies on complex internal representations that are learned and better specified gradually with each exposure or recall.

Participants in the gradual feedback condition, who were more willing to commit errors, showed more gradual learning between blocks than participants who received binary feedback as seen in the significantly lower maximum trial-to-trial improvement. It is not possible to discern from our design whether this means they learned more gradually, or whether we simply gained more insight into their internal process even though the underlying process is equally gradual. It is possible that the process of explicitly rewarding partial answers and focusing on partial overlap during feedback could lead learners to adopt a different learning strategy that results in more gradual learning than usual. We surmise, however, that this difference is largely driven by the latter, meaning that the manipulation exposes information about incomplete latent representations that are present in all learners but usually hidden.

As information about the learner's progress with individual words is relevant for intelligent tutoring systems, stimulating learners to commit mistakes helps to gain an insight into these granular steps (Amaral & Meurers, 2007; Cook & Payne, 2002). Levenshtein distance proves a useful tool for measuring the overall distance in similarity from the learner's representation to the target word. Our study employed a very simplistic mechanism to address the representational distance between error and target word in order to maintain parity between the experimental conditions. For educational purposes, learning tools could go further and address specifically the parts of the word that were not learned correctly, for example, by highlighting letter overlap, or contrasting the error with other similar words with which the user might have confused the correct answer.

5.4 Insights for Learning and Teaching

Perhaps the most straightforwardly applicable outcome of this research for practice lies outside the three core hypotheses, namely in the motivational boost of gradual feedback reported by participants. The more positive feedback to partially correct answers was mentioned in the post-experiment questionnaire as 'encouraging' and by some as 'useful for reflection', resulting in a better enjoyment of the learning task. Although this is true for most participants, some found the partial feedback more confusing, so a personalized approach is probably best. An increase in motivation likely results in more attention and higher willingness to extend cognitive effort on word learning and thus improves learning outcomes in the short and long term.

Gradual feedback is easy to implement in digital learning applications. Where most teachers in person-to-person language learning environments are probably already intuitively inclined to praise nearly-correct answers and to point out exactly where the learner went wrong, hopefully this research can help computer-assisted learning to provide the same benefit. Error generation increases insight into the learners internal understanding and thus allows for a more targeted education method.

An increase in enjoyment of learning tasks is a goal of gamification, which is increasingly common in learning applications (Dehghanzadeh et al., 2021). However, many extrinsic motivators used such as points, badges, and leaderboards often work only in the short term and see a decline in effect over time (Toda et al., 2018). Effective and integrated design of such devices is required to achieve long-term gains. Similarity-based feedback can easily be integrated in gamification applications and allows for all-important alignment of educational and game mechanics (Lim et al., 2015). Our finding of possible habit formation based on instructions and type of feedback might represent a more long-lasting effect than an otherwise simple point system. A similar distance-based feedback mechanic can also easily be translated to other domains where a similarity measure can be defined.

6 CONCLUSION

Gradual feedback participants committed answers much more frequently than control during learning, immediate post-test, and even in the one-week delayed post-test. This increase in error-commission may have also led to overall learning gains, though we could not confirm a significant main effect of the gradual condition. Gradual feedback participants did better over time, suggesting that the gradual feedback helped better consolidate orthographic representations. Across both conditions, word learning was shown to be gradual, in that participants often learn parts of the word in several steps before having fully learned the word. With the increased error commission due to gradual feedback, this gradual shaping of the word representation can be followed online, allowing for more tailored feedback

in both digital and non-digital vocabulary learning. The presentation of gradual feedback was also perceived as motivating and improved enjoyment of the learning task. Hence, gradual Levenshtein-based feedback promises to be a useful addition to digital word learning applications.

ACKNOWLEDGEMENTS

We would like to thank Peta Baxter, Randi Goertz, and Josh Ring for their feedback and help in designing this study.

This project was funded by the Netherlands Initiative for Education Research of the Dutch Research council under grant 405.17300.048.

REFERENCES

Abraham, D., McRae, K., & Mangels, J. A. (2019). "A" for effort: Rewarding effortful retrieval attempts improves learning from general knowledge errors in women. *Frontiers in Psychology, 10*, 1179.

Amaral, L., & Meurers, D. (2007). Conceptualizing student models for ICALL. In C. Conati, K. McCoy, & G. Paliouras (Eds.). *User Modeling 2007: 11th International Conference, UM 2007, Corfu, Greece, July 25-29, 2007. Proceedings 11* (pp. 340-344). Berlin / Heidelberg: Springer.

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language, 73*, 116–130.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. In arXiv [stat.CO]. arXiv.

Baxter, P., Droop, M., van den Hurk, M., Bekkering, H., Dijkstra, T., & Leoné, F. (2021). Contrasting similar words facilitates second language vocabulary learning in children by sharpening lexical representations. *Frontiers in Psychology, 12*, 688160.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *23*(3), 321–343.

Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition, 25(2)*, 185–201.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109(2)*, 204–223.

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2022). The most appropriate lexical unit for L2 vocabulary research and pedagogy: a brief review of the evidence. *Applied Linguistics, 43(3)*, 596–602. Butler,

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review, 18(6)*, 1238–1244.

Butterfield, B., & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. Brain Research. *Cognitive Brain Research, 17(3)*, 793–817.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. *In F. Mosteller & R. F. Boruch (Eds.), Evidence matters: Randomized trials in education research* (p. 174). Brookings Institution Press.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364(1536)*, 3773–3800.

Daw, N. D., & Courville, A. C. (2008). The pigeon as particle filter. *Advances in Neural Information Processing Systems 20*, 369–376.

De Grauwe, S., Willems, R. M., Rueschemeyer, S.-A., Lemhöfer, K., & Schriefers, H. (2014). Embodied language in first- and second-language speakers: neural correlates of processing motor verbs. *Neuropsychologia, 56*, 334–349.

De Groot, A. M. B. (1995). Determinants of bilingual lexicosemantic organization. *Computer Assisted Language Learning, 8(2-3)*, 151–180.

Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2021). Using gamification to support learning English as a second language: a systematic review. *Computer Assisted Language Learning, 34(7)*, 934–957.

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism, 5(3)*, 175–197.

Filsecker, M., & Hickey, D. T. (2014). A multilevel analysis of the effects of external rewards on elementary students' motivation, engagement and learning in an educational game. *Computers & Education, 75*, 136–148.

Gooch, D., Vasalou, A., Benton, L., & Khaled, R. (2016). Using gamification to motivate students with dyslexia. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 969–980.

Groot, P. J. M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology*.

Heimbeck, D., Frese, M., Sonnentag, S., & Keith, N. (2003). Integrating errors into the training process: the function of error management instructions and the role of goal orientation. *Personnel Psychology, 56(2)*, 333–361.

Hooshyar, D., Pedaste, M., Saks, K., Leijen, Ä., Bardone, E., & Wang, M. (2020). Open learner models in supporting self-regulated learning in higher education: A systematic literature review. *Computers & Education, 154*, 103878.

Janssen, C., Segers, E., McQueen, J. M., & Verhoeven, L. (2015). Lexical specificity training effects in second language learners. *Language Learning, 65(2)*, 358–389.

Karpicke, J. D. (2017). 2.27 - Retrieval-based learning: a decade of progress. In J. H. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference (Second Edition)* (pp. 487–514). Academic Press.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35(4)*, 989–998.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 41(1)*, 283–294.

Lally, P., & Gardner, B. (2013). Promoting habit formation. Health Psychology Review, 7(sup1), S137–S158.

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

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54(3)*, 399–436.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: effects of language learning context. *Language Learning, 48(3)*, 365–391.

Levenshtein, & Others. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

Lim, T., Carvalho, M. B., Bellotti, F., Arnab, S., de Freitas, S., Louchart, S., Suttie, N., Berta, R., & De Gloria, A. (2015). The LM-GM framework for serious games Analysis. Citeseer.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*, 465–489.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092.

Moridis, C. N., & Economides, A. A. (2008). Toward computer-aided affective learning systems: a literature review. *Journal of Educational Computing Research, 39(4)*, 313–337.

Palmberg, R. (1987). Patterns of Vocabulary Development in Foreign-Language Learners1. *Studies in Second Language Acquisition, 9(2)*, 201–219.

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(6)*, 1023

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). Latent growth curve modeling. SAGE.

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism , 15(1)*, 157–166.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1848–1858.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research: JSLHR*, *49*(6), 1175–1192.

Talmi, D. (2013). Enhanced Emotional Memory: Cognitive and Neural Mechanisms. *Current Directions in Psychological Science, 22(6)*, 430–436.

Toda, A., Valle, P. H. D., & Isotani, S. (2018). The dark side of gamification: an overview of negative effects of gamification in education. In A. I. Cristea, I. I. Bittencourt, & F. Lima (Eds.), *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions: First International Workshop on Social, Semantic, Adaptive and Gamification Techniques and Technologies for Distance Learning, HEFA 2017, Maceió, Brazil, March 20–24, 2017, Revised Selected Papers 1* (pp. 143–156). Springer International Publishing.

Tulis, M., & Ainley, M. (2011). Interest, enjoyment and pride after failure experiences? Predictors of students' state-emotions after success and failure during learning in mathematics. *Educational Psychologist, 31(7)*, 779–807.

van Heuven, W. J. B., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language, 39(3)*, 458–483.

Wilson, B. A., Baddeley, A., Evans, J., & Shiel, A. (1994). Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological Rehabilitation, 4(3)*, 307–326.

Woods, D. L., Kishiyamaa, M. M., Lund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F., & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology, 33(1)*, 101–111.

Young, D. J. (1991). Creating a low-anxiety classroom environment: What does language anxiety research suggest? *The Modern Language Journal, 75(4)*, 426–439.

APPENDIX

All stimuli and data used in this experiment, the code used for data collection and analysis, as well as more detailed reports of the generalized linear models used in this paper can be found on the Donders Repository: https://doi.org/10.34973/zz4g-ma56