

Variation in Learning Outcomes Between Gamified and Traditional Learning Methods

Anthony Faldetta, Isaac Gurney, Carmen Reyes, University of Washington

Email: antman10@uw.edu, isaacg6@uw.edu, creyes6@uw.edu

Abstract: This proposal examines whether a gamified language learning platform, Duolingo, improves immediate German vocabulary test performance and self-efficacy compared with a traditional flashcard approach among students at the University of Washington. In the proposed study, participants will study a matched set of 20 German words and phrases delivered either through Duolingo or through flashcards constructed from the same content, then complete a short online assessment. The assessment includes an eight-item graded vocabulary score (vocab_total, 0 to 8) and seven self-efficacy statements rated on a five-point agreement scale and averaged into a self-efficacy index (se_index). A small pilot using this design shows no statistically significant differences between Duolingo and flashcards on vocab_total ($p \approx .50$, one-tailed t) or se_index ($p \approx .22$). These pilot results illustrate the planned analysis pipeline and suggest that, under tightly matched content in a single short study session, the two methods may perform similarly on both immediate recall and perceived capability.

Introduction

Popular gamified learning platforms such as Duolingo, Kahoot, and Khan Academy are ubiquitous in the modern educational system. They entice users by teaching lessons using streaks, badges, leaderboards, and points, raising a rational question: Do these features translate into genuine learning and comprehension, or do they merely channel users' engagement? Our group pursued this topic after discovering our shared personal experiences with Duolingo's lessons that "did not really stick," prompting a focus on an individual's motivation to keep studying and information retention as dual outcomes worth measuring.

The problem we address in our study is whether or not gamified learning actually helps students learn meaningfully, rather than simply boosting engagement through rewards and streaks. Many gamified platforms are designed around points, badges, and leaderboards that optimize for clicks and daily use rather than building deep conceptual understanding, giving learners an artificial sense of progress that does not always match their ability to recall or apply what they have learned. Due to this, prior research on gamified language learning and self-efficacy highlights the need to measure both immediate test performance and learners'

confidence in their understanding, to ensure that the enjoyment of the game is not mistaken for genuine learning.

To accomplish this task, we designed a controlled, head-to-head comparison in which college students at the University of Washington would study the same German words and phrases using either a gamified learning app (Duolingo) or via flashcards prepared from the same materials. Then, the students would take the same online assessment and brief self-efficacy survey. This design deliberately aligns content and exposure so that any performance difference can be attributed to the learning method rather than coverage, ensuring the internal validity of the study.

Our working research question is: *To what extent do gamified learning systems vary in their ability to help college students learn German vocabulary compared to traditional learning methods, and how do they affect motivation and self-efficacy?*

To make the proposal concrete, we also articulated two focused sub-questions that align with our measures:

1. Does Duolingo improve immediate German vocabulary test scores relative to flashcards?
2. Does Duolingo improve self-efficacy relative to flashcards?

Contributions of this work are (a) a matched-content design for a direct comparison of gamified vs. traditional learning on immediate retention of vocabulary terms, (b) a clear measure of perceived learning through a seven-item self-efficacy index, and (c) an analysis pipeline (cleaning, index construction, and one-tailed t-tests) that can be replicated and scaled.

Literature review

Gamification, mobile-assisted language learning, and instructional design. Language learning apps, such as Duolingo, popularize game elements, including points, streaks, and

leaderboards, to sustain engagement, but whether these mechanics actually improve users' learning depends on how they are designed and measured. A systematic review of Duolingo's curriculum maps a mixed-evidence base that spans various methods, contexts, and outcomes, making it difficult to isolate the effects of gamification without tighter controls (Shortt, Tilak, Kuznetcova, Martens, & Akinkuolie, 2021). Comparative work outside of language learning shows that when game elements are embedded directly into instructional activities, they can drive achievement: in Grade 9 mathematics, flipped-learning with gamification conditions significantly outperformed both traditional and online-independent formats on test scores and cognitive engagement (Lo & Hew, 2020). Together, these findings motivate content-matched, head-to-head studies that compare gamified tools with traditional baselines using clearly defined outcome measures.

Motivation, self-regulation, and self-efficacy as co-outcomes. Gamified systems often target motivational pathways, such as autonomy, competence, and feedback, that are highlighted in self-regulated learning (SRL) and self-determination theory (SDT). Evidence links SRL and SDT-aligned support to persistence and strategic action in learning tasks (Brenner, 2022). Recent work also shows that self-efficacy can positively moderate the impact of educational experiences on downstream intentions and behaviors, suggesting that learners who feel more capable are more likely to persist and apply what they have learned (Pham, Quoc Lam, & Tran Le, 2023). For language learning tasks, this supports treating motivation and perceived capability as co-outcomes with performance, and justifies measuring not only recall but also perceived capability through a self-efficacy index.

Language technology, design features, and content presentation. Recent studies on gamified conversational agents in language classes apply relatively simple design methods, such as setting

clear goals and showing progress indicators. These features tend to boost learners' motivation, which is useful if the goal is to keep students practicing over time (Khosrawi-Rad et al., 2025). However, these studies rarely run direct A/B tests against traditional methods that use the same content and an immediate vocabulary assessment, so they can show that motivation increases without settling whether vocabulary recall is actually better than with a non-gamified study. At the same time, even when two groups study the same vocabulary items, apps and study sheets can present those items in different ways, for example, through example sentences, pronunciation guides, or part-of-speech labels. Work in specialized lexicography shows that such presentation choices can change how well people remember items (Sviķe, 2021). This means a study could accidentally credit gamification for gains that really come from how the information is displayed, unless word lists and basic supports are kept comparable across conditions.

Gap and rationale. Two research gaps in this evidence base matter for our study: very few designs directly compare a gamified learning tool and a traditional learning method on the same items with an immediate test, and even fewer track both test performance and self-efficacy simultaneously. Our proposal responds to these gaps by using a 20-item café-themed German lexicon, an 8-item graded assessment drawn from that lexicon, and a 7-item self-efficacy index. Keeping content coverage and basic presentation support comparable across conditions allows us to more cleanly test whether differences in immediate vocabulary performance and self-efficacy stem from the gamified delivery format or the traditional flashcard approach.

Methodology

Proposed Data Collection

Context and participants. The proposed study focuses on college students at the University of Washington who volunteer to participate. Recruitment will occur through informal outreach

such as word of mouth and course-related announcements. Participants complete the survey at home on their own devices. Those assigned to the gamified condition use the Duolingo app or website on a smartphone or computer. Those assigned to the traditional condition use a digital or printed flashcard set that is constructed from the same vocabulary list. The study is intended for students who are not already fluent in German.

Learning materials (matched). The German café vocabulary set defines the taught content. It contains 20 words and short phrases related to ordering food and drinks. The Duolingo condition covers these items through standard Duolingo lessons. The traditional condition uses flashcards built from the same German words and English translations, with no additional embellishments, so that content coverage is matched as closely as possible. We began by completing the first set of Duolingo lessons to discover which concepts were covered. We then used the core concepts and vocabulary from the first Duolingo lessons to create the flashcards. The flashcards were made using Google Slides so participants could easily go through our materials and flip between English and German translations of the same phrase. We used ChatGPT to generate phonetic spellings and pronunciations for each of the German words or phrases. This was done in an attempt to match some of the vocal and audible elements of Duolingo to mitigate the potential unfairness of having an application where you can hear someone speaking the words. Not to say Duolingo is better because of this, we simply wanted to even the playing field to create the most controlled study possible. Once we had our study materials fully fleshed out, we could begin recruiting participants, assigning them to one of the treatment groups, and completing our quiz and reflection. This is discussed in more depth throughout the following “Procedure” section.

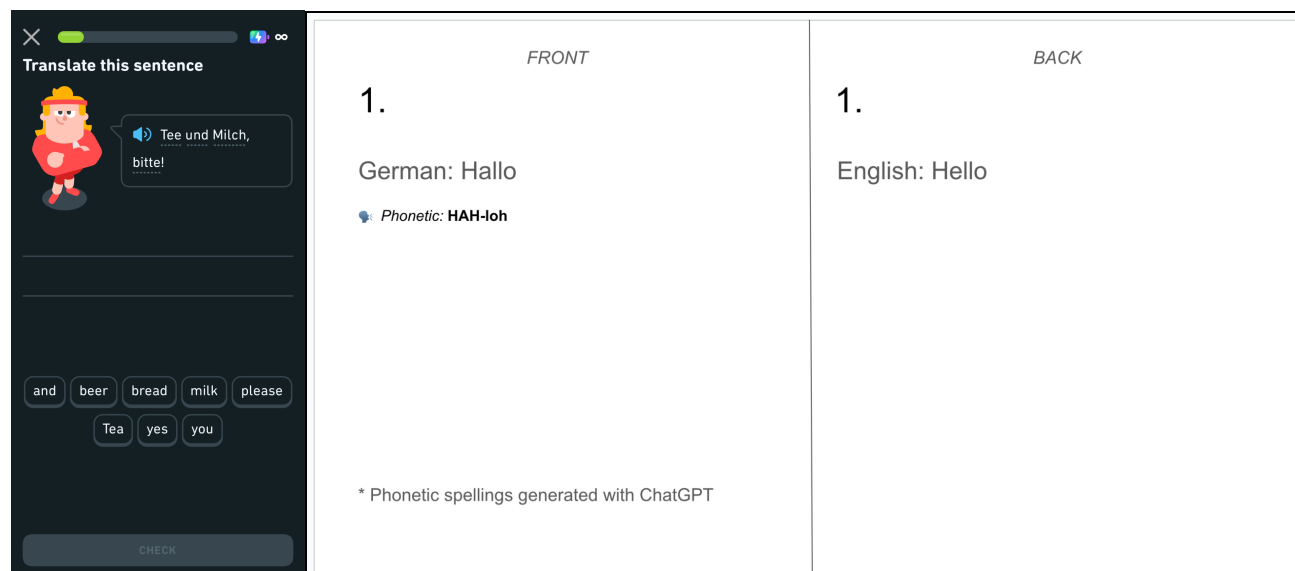


Figure 1. Study Materials: Duolingo vs. Flashcards Example

Procedure

1. Randomly assign each participant to either the Duolingo condition or the flashcard condition.
2. Provide study instructions and access to the assigned materials, and allow a short learning period of 15 minutes during which participants study independently.
3. Immediately after the study window, administer the online vocabulary assessment followed by a self-efficacy survey.

This procedure is designed so that both groups have the same total study time and the same vocabulary coverage, differing only in whether the material is delivered in a gamified or traditional format.

Measures

- *Vocabulary performance (primary outcome).* vocab_total (0 to 8) is an eight-item graded assessment. Each item is scored 1 for correct and 0 for incorrect, and the items are

summed. Although the taught set includes 20 words and phrases, only eight items are graded to produce the test outcome.

- *Self-efficacy (secondary outcome)*. *se_index* is constructed from seven statements about perceived learning and motivation, each rated 1 to 5 from Strongly Disagree to Strongly Agree. The seven items are averaged to create a single index for each participant.

Data preparation

Survey responses for vocabulary are entered as text. Each response is coded as correct or incorrect (1 or 0) and summed to compute *vocab_total*. The self-efficacy/efficacy ratings are converted to numeric values from 1 to 5 (1 being strongly disagree and 5 being strongly agree) and averaged to compute *se_index*. We standardized small differences in how missing values were written, such as “NA” versus “na,” so that all missing entries used the same code in the dataset. To protect privacy, names and emails were removed before any analysis. The data preparation steps are intentionally simple and reproducible so that other researchers can follow the same pipeline and verify results.

Hypothesis

- **Learning (*vocab_total*).** **H₀:** There is no difference between Duolingo and flashcards in the scores of our German language assessment.
H₁ (one-tailed): Duolingo yields higher test scores than flashcards.
- **Self-efficacy (*se_index*).** **H₀:** Duolingo will show no improvement to participants' self-efficacy compared with traditional learning methods.
H₁ (one-tailed): Duolingo yields higher self-efficacy than flashcards.

Statistical tests. Report descriptive statistics for both outcomes, then run independent-sample

one-tailed t-tests with $\alpha = 0.05$ to compare Duolingo vs. flashcards on vocab_total and se_index. Optionally, assess the correlation between the two outcomes.

Results

We summarize vocab_total and se_index with means, standard deviations, medians, modes, ranges, and counts. In the pilot, vocab_total averaged 6.76/8 with $SD \approx 0.83$, median = 7, mode = 7, range = 3 (min 5, max 8), $n = 21$. se_index averaged ≈ 3.97 on the 1–5 scale with $SD \approx 0.54$, median = 4, mode = 4.5, range = 2 (min 2.83, max 4.83), $n = 21$. These values characterize the overall sample; they are not yet stratified by condition. Below are tables of our results.

se_index	se_index	vocab_total	vocab_total
Mean	3.96825397	Mean	6.76190476
Standard Error	0.11875626	Standard Error	0.18132793
Median	4	Median	7
Mode	4.5	Mode	7
Standard Devi	0.54420954	Standard Devi	0.83094897
Sample Variar	0.29616402	Sample Variar	0.69047619
Kurtosis	-0.4538348	Kurtosis	-0.4736467
Skewness	-0.4127285	Skewness	-0.0821224
Range	2	Range	3
Minimum	2.83333333	Minimum	5
Maximum	4.83333333	Maximum	8
Sum	83.3333333	Sum	142
Count	21	Count	21

Table 1: Descriptive statistics

t-Test: Two-Sample Assuming Unequal Variances		
	<i>duolingo vocab_total</i>	<i>flashcards vocab_total</i>
Mean	7	6.4
Variance	0.727272727	0.488888889
Observations	12	10
Hypothesized Mean Differenc	0.6	
df	20	
t Stat	-1.00655E-15	
P(T<=t) one-tail	0.5	
t Critical one-tail	1.724718243	
P(T<=t) two-tail	1	
t Critical two-tail	2.085963447	

Table 2: T-test - Does Duolingo outperform flashcards?

t-Test: Two-Sample Assuming Unequal Variances		
	<i>duolingo se_index</i>	<i>flashcards se_index</i>
Mean	3.944444444	3.933333333
Variance	0.35016835	0.272839506
Observations	12	10
Hypothesized Mean Differer	0.2	
df	20	
t Stat	-0.794910265	
P(T<=t) one-tail	0.217999017	
t Critical one-tail	1.724718243	
P(T<=t) two-tail	0.435998033	
t Critical two-tail	2.085963447	

Table 3: T-test - Does Duolingo lead to greater self-efficacy than traditional flashcards?

Primary tests. To test the directional hypotheses, we use independent-samples, one-tailed t-tests ($\alpha = .05$) to compare Duolingo vs. flashcards on each outcome. Given realistic differences in variance between groups, we treat the unequal-variance version (Welch) as appropriate. The decision rule is standard: if the one-tailed p-value is below .05 in the predicted direction, reject H_0 in favor of H_1 ; otherwise, retain H_0 .

Correlation (descriptive). We compute a simple correlation between *vocab_total* and *se_index* to describe whether higher immediate recall tends to travel with higher perceived capability in this sample. This is not a confirmatory test, but it helps characterize the relationship between the two outcomes.

Pilot results. In the pilot data, the one-tailed comparison for vocab_total yields $p \approx .50$; for se_index, $p \approx .22$. Both tests fail to cross the $\alpha = .05$ threshold in the hypothesized direction, so we retain the null hypotheses for both outcomes. Under matched content and immediate testing, the pilot does not support the claim that Duolingo outperforms flashcards on either immediate recall or perceived capability.

Discussion (anticipated)

Implications for Research Question. To reiterate, our discussion was “*To what extent do gamified learning systems vary in their ability to help college students learn German vocabulary compared to traditional learning methods, and how do they affect motivation and self-efficacy?*” Overall, our results showed no indication that gamified education has any impact on learning outcomes, both externalized and internalized, in University of Washington students.

What the pilot suggests. The pilot pattern is straightforward: no detectable difference between Duolingo and flashcards on the immediate vocabulary test or on self-efficacy at the current sample size. Within this design, that means either (a) the two methods are roughly equivalent for short, fixed-content vocabulary study, or (b) there is a difference, but the study is underpowered to detect it. Both interpretations are compatible with the data; the design choices we made help ensure that if a difference exists, it would not be because one group studied more items or saw richer lexical supports.

Why immediate tests matter here. Since the outcome is an immediate post-study assessment, the comparison focuses on what learners can recall right after a short study window. Many gamified systems are built to support spaced repetition and habit formation. Those strengths might appear with longer exposure or delayed tests, but they are not being evaluated here by

design. Immediate testing removes time as a confound and makes the comparison clean in a single session of study.

Placing the result alongside prior work. Findings that engagement can rise without clear score gains are consistent with a broad picture in the literature: design structure matters (e.g., flipped plus gamification), and measurement matters (what is tested, when it is tested). Reviews of Duolingo studies emphasize that across many contexts and measures, it is difficult to generalize about learning outcomes without tight control of content and timing. By pairing a short recall test with a self-efficacy index, this project ensures that any benefit of gamification that is primarily effective (e.g., “I feel more capable”) would be visible, even if scores stay flat.

Implications for learners and designers. If further data replicate the pilot pattern, one implication is that for a brief, fixed-content vocabulary task, a simple non-gamified method (flashcards built from the same content) can perform as well as a gamified app on both measured outcomes. That does not argue against gamified platforms broadly. Rather, it suggests that when time is short and content is fixed, the added game mechanics may not change immediate recall or immediate self-efficacy compared with a straightforward study tool. For tool designers, that indicates improving clarity of presentation, ease of practice, and consistency of coverage should be considered just as critical as incorporating game elements for this use case.

What scaling and extensions could show. The next step is to increase the sample size while keeping the content matched. A larger (n) strengthens inference and allows checks for group imbalances. A planned delayed retention checkpoint (e.g., 24–72 hours) would test whether the gamified condition helps learners remember material after time passes. If differences emerge with longer exposure or spacing, that pattern would naturally point to time-on-task and repetition as the mechanism, without changing the core conclusion here about immediate recall under

matched conditions. Throughout, the content-matching principle should remain in place so that any observed changes can be attributed to the method rather than to coverage.

Conclusion

Pilot results from our small sample suggest that there may be no large immediate difference in objective German café vocabulary test scores between a gamified learning platform (Duolingo) and traditional flashcards. Likewise, we did not observe a clear difference in self-efficacy between the two conditions. Within the limits of this pilot, both approaches appear to support short-term vocabulary learning and perceived capability to a similar extent.

However, these findings are necessarily tentative. Our pilot used a limited set of vocabulary items and a single short study interval, which does not reflect Duolingo's core design goal of building habits and spaced repetition over time. In the proposed full study, we will extend the study period and ensure that each vocabulary term receives a comparable number of practice repetitions in both conditions. This will provide a more realistic test of whether gamified features yield advantages for retention or motivation beyond what can be achieved with carefully designed flashcards.

Sampling is another important limitation. Because pilot participants were recruited through convenience sampling from peers and close friends, the results are not representative of the broader population of University of Washington students. In the full study, we plan to use either simple random sampling from a larger pool or a matched-pairs design in which participants are paired based on pre-test vocabulary scores and randomly assigned within each pair to Duolingo or flashcards.

From an ethical perspective, the proposed study is straightforward and low-risk. We will obtain informed consent, clearly describe what demographic information is collected and how it

will be used, and ensure that no personally identifying information is reported in any publication or shared dataset. All survey and test data will be stored securely and de-identified before analysis. With these safeguards in place, the proposed research can be carried out ethically while providing useful evidence about when gamified learning platforms offer advantages over traditional methods.

References

- Brenner, C. A. (2022). Self-regulated learning, self-determination theory and teacher candidates' development of competency-based teaching practices. *Smart Learning Environments*, 9, Article 3. <https://doi.org/10.1186/s40561-021-00184-5>
- Khosrawi-Rad, B., Keller, P. F., Benner, D., Grogorick, L., Borchers, A., Janson, A., Leimeister, J. M., & Robra-Bissantz, S. (2025). Promoting students' motivation in language education with gamified pedagogical conversational agents. *Computers & Education*, 238, 105374. <https://doi.org/10.1016/j.compedu.2025.105374>
- Lo, C. K., & Hew, K. F. (2020). A comparison of flipped learning with gamification, traditional learning, and online independent study: The effects on students' mathematics achievement and cognitive engagement. *Interactive Learning Environments*, 28(4), 464–481. <https://doi.org/10.1080/10494820.2018.1541910>
- Pham, M., Quoc Lam, B., & Tran Le, V. P. (2023). The e-entrepreneurial intention of students: The role of self-efficacy and education. *Entrepreneurial Business and Economics Review*, 11(1), 127–143. <https://doi.org/10.15678/EBER.2023.110107>
- Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2021). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning*. Advance online publication. <https://doi.org/10.1080/09588221.2021.1933540>
- Sviķe, S. (2021). Mobile apps as language-learning tools: Challenges, problems and solutions of specialised lexicography. *AILA Review*, 34(1), 19–36. <https://doi.org/10.1075/aila.20006.svi>