Online Journal of Communication and Media Technologies, 2024, **14**(2), e202426
e-ISSN: 1986-3497

OPEN ACCESS

Research Article

The advantages of gamification for collecting linguistic data: A case study using Word Ladders

Francesca Genovese [1]
0009-0000-7468-5987

Marianna Marcella Bolognesi [1*]
0000-0002-3292-8968

Angelo Di Iorio [1]
0000-0002-6893-7452

Fabio Vitali [1]
0000-0002-7562-5203

[1] University of Bologna, Bologna, ITALY
[*] Corresponding author: m.bolognesi@unibo.it

Citation: Genovese, F., Bolognesi, M. M., Di Iorio, A., & Vitali, F. (2024). The advantages of gamification for collecting linguistic data: A case study using Word Ladders. *Online Journal of Communication and Media Technologies, 14*(2), e202426. https://doi.org/10.30935/ojcmt/14443

ARTICLE INFO

Received: 20 Sep 2023

Accepted: 4 Mar 2024

ABSTRACT

This paper delves into the integration of gamification techniques within the field of linguistics to enhance data collection for academic research purposes. Through an exploration of the Word Ladders mobile application, designed to elicit hierarchical word associations and therefore linguistic data, the study investigates the potential benefits of gamification in terms of data quality, user experience, and motivation in taking part to the research and to the data collection task. The experimental design examines the advantages of a gamified approach compared to traditional research methods (online surveys), through an experimental session followed by a survey (n=189). Results showed that competition between users is a powerful motivator that can be easily integrated in gamified approaches and less so in classic online surveys, driving engagement and potentially enhancing the scalability of data collection while retaining the quality of data collected in classic lab settings. While challenges persist, our research contributes to the understanding of gamification's impact on data collection, user experience, and motivation, laying the foundation for transformative advancements in the field of language and communication sciences.

Keywords: gamification, linguistics, cognition, motivation, usability, user-experience

INTRODUCTION

In recent years, gamification techniques have seen a growing presence in educational applications (Al Saad and Durugbo, 2021; Hawamdeh and Soykan, 2021; Ishaq et al., 2021, Sever et al., 2015; Zheltukhina et al., 2023), although the development of gamified tools for the collection of linguistic data to be used in academic contexts remains relatively uncommon.

An initial conceptualization of the term "gamification" emerged in 2008, as documented in a blog post authored by Brett Terill, CEO and co-founder of Tenuki, Inc., a company specializing in the development of social games (cited in Welbers et al., 2019). Brett Terill's definition pertained to the utilization of game mechanics in various online contexts to enhance user engagement.

Deterding et al. (2011) later proposed a more widely accepted definition in which gamification is the integration of game design elements into non-game settings. Therefore, while games primarily aim to provide entertainment, gamification employs game elements to motivate users to engage with unprecedented intensity and duration in the fulfilment of various types of tasks (Flatla et al., 2011; Zichermann & Cunningham, 2011).

A related notion is that of "serious games", which are defined as "interactive computer-based games developed with the purpose of offering more than mere entertainment" (Ritterfeld et al., 2009). The primary goal of serious games therefore is to attain specific outcomes for the gamer, such as enhancing learning, skill development, increasing awareness, or influencing behavioral change. These games are intentionally crafted to tackle real-world challenges and offer users an immersive and interactive environment to acquire and apply new knowledge or abilities. By simulating real-life scenarios, incorporating problem-solving tasks, and incorporating feedback and assessment mechanisms, serious games enable users to actively engage, measure progress, and evaluate performance.

In many academic scenarios, the tasks assigned to test users can be challenging, monotonous, and repetitive, leading to participant disinterest and disengagement. In these contexts, gamification offers a potential solution by motivating individuals to participate in experiments through the implementation of feedback mechanisms, user competition, rewards, and other relevant gamification elements as recently argued by Long et al. (2023).

It is important, however, to ensure that the introduction of gamification does not compromise the quality of the data collected, thereby maintaining its comparability to data obtained through traditional research methods (Germine et al., 2012).

As a matter of fact, the integration of gamification techniques in the fields of language and communication sciences has not yet gained widespread adoption, despite its well-documented efficacy in enhancing engagement, motivation, and learning outcomes across diverse disciplines. Notably, recent research conducted by Short et al. (2021) and Smith et al. (2022) suggest that the integration of gamification in language and communication studies holds promising potential for improving engagement, proficiency, and interaction in language learning and communication processes. As scholars and practitioners become increasingly aware of these advantages, we anticipate a rise in exploration and implementation of gamification techniques within these domains, leading to transformative advancements in language and communication sciences.

As mentioned, gamification in educational contexts and for linguistic analysis has been discussed previously (in particular AlSaad & Durugbo, 2021; Hawamdeh & Soykan, 2021). Yet, while gamification has been extensively employed to enhance learning processes, its application as a means of gathering high-quality and large-scale linguistic data for academic purposes has yet to gain widespread traction. Consequently, the untapped potential of gamification in facilitating data collection for quantitative linguistic analyses remains a subject for further investigation and exploitation by researchers in the field.

Considering this context, we have undertaken the testing of gamification in the field of linguistics to explore the potential benefits, particularly in terms of user behavior, while collecting data for academic purposes.

Our hypotheses revolve around investigating the potential benefits of gamification techniques in data collection in the linguistic domain, specifically focusing on three research hypotheses.

H1. **On data quality**. Overall, we expect that data obtained through gamification techniques will exhibit (at least) the same quality as data collected through traditional methods, such as classic survey forms. Data quality is measured directly using two metrics: linguistic productivity (the number of linguistic data produced) and accuracy (the relevance of the linguistic data produced).

H2. **On user experience**. We expect to find that a gamified approach to language tasks enhances users' motivation, engagement, and perceived ease in performing the task. We expect this benefit to be particularly strong for the group of participants that played in 'challenge' mode–one against each other–compared to the group of participants who played in the "individual" mode, as motivated in the section of the current paper called "theoretical background".

H3. **On users' motivation**. We expect to find that users who engage in the challenge mode, competing against other real individuals, including acquaintances, will exhibit higher motivation to perform the tasks and engage in future gameplay, compared to participants who used the individual mode or performed the task on a standard online survey.

Through our research, we aim to examine these hypotheses and shed light on the benefits of gamification techniques in data collection, including their impact on data quality, user experience, and sustained motivation for continued engagement.

THEORETICAL BACKGROUND

To Gamify or Not to Gamify, This is the Question

Experimental research in academia involves very often tasks that participants consider challenging, boring, or repetitive. This can cause participants to become disinterested and disengaged, or to feel particularly stressed because of the experimental setting. These factors can arguably have a negative impact on the quality of the data collected through traditional experimental tools (McPherson & Burns, 2007).

In this case, the implementation of gamification as a strategic approach proves to be a promising solution that significantly enhances participants' engagement levels, improves data quality, and mitigates experiential constraints. Gamification has been described as "the use of game-based mechanics and game-based design elements in non-game settings to engage users and encourage achievement of desired outcomes through motivation of users" (Reiners et al., 2015). As it appears, increasing and supporting the *motivation of participants* is the main justification of applying game-based approaches to non-game applications, and such increase needs to be identified as the main applicable metrics (e.g., in addition to Reiners et al., 2015, see Deterding et al., 2011 and Sailer et al., 2023).

Specifically in the linguistic domain, a sizable number of gamification experiences have been described in the scientific literature, ranging from the use of game-like mechanics for second-language acquisition (too many to list, and even systematic reviews abound, see for instance Benini & Thomas, 2021; Dehghanzadeh et al., 2021; Short et al., 2023; Waluyo et al., 2023), for grammar proficiency of one's own first language (e.g., Eryigit et al., 2021; Krisbiantoro, 2020; Purgina et al., 2020) to collection of data, visualization and explanation of advanced linguistic issues (e.g., Bonetti & Tonelli, 2021; Eryigit et al., 2023; Kim et al., 2023; Ogawa et al., 2020; Sevastjanova et al., 2021).

However, it is worth noting that the research findings on the benefits of gamified techniques in academia are not consistently positive (Lumsden et al., 2016). In fact, behavioral research utilizing classic online surveys and crowdsourcing tasks may yield more accurate data under certain conditions, potentially attributed to the context in which the data is collected. Participants, aware of the institutional setting, may feel compelled to pay greater attention to the assigned task, thereby generating higher-quality data in the lab, compared to the gamified setting. Therefore, two primary factors determine whether a gamified approach is preferable over a more conventional academic approach. The first factor pertains to the user population targeted for data collection, while the second factor concerns the nature of the task and its implementation through gamification. Regarding the first factor, research shows that gamified tasks may work particularly well with participants with certain clinical conditions. Lumsden et al. (2016) for instance review six studies that provide empirical evidence in support to the claim that performance problems in ADHD patients may be alleviated by gamified settings, because this population is particularly sensitive to quick rewards and immediate feedback, which are easily implemented in games. Even within the population of healthy users, it has been shown that individual characteristics affect the potential advantage of a gamified over a non-gamified approach to learning in educational settings, for instance (Smiderle et al., 2020). The second factor deals with the type of task. In fact, not all cognitive tasks equally show benefits of a gamified approach compared to traditional experimental settings like online surveys. Gamified approaches have shown advantages over classic experimental settings in various tasks (Hammady & Arnab, 2022 for a recent systematic literature review). For instance, gamification shows advantages over classic methods in training simulations aimed at changing users' behavior toward health-related issues (e.g., Verduin et al., 2013; Wright & Bogost, 2007) or societal problems, such as raising sensitivity to environmental issues (Ro et al., 2017). Gamified approaches extensively show advantages also in educational programs (Qian & Clark, 2016; Smiderle et al., 2020). While gamification can be effective in many cognitive and linguistic tasks, there are tasks, where gamification may not provide significant benefits or may even be counterproductive. For instance, in tasks that involve straightforward data collection, standardized assessments, or controlled experimental paradigms, the addition of game elements might not significantly enhance participant performance or engagement. Conversely, it may introduce uncontrolled variables that may be conflated with the experimental variables (Gundry & Deterdring, 2019). Also, when a given task is based on a simple decision, like tasks based on simple cognitive processes (basic perception, recognition, or basic language comprehension) the data might not benefit significantly from game elements.

The Present Study: Introducing Word Ladders

Word Ladders, a mobile application accessible on Apple and Android platforms, has been developed by the Abstraction research group as part of a five-year project funded by the European Research Council (grant agreement: ERC-2021-STG-101039777). Functioning as a linguistic game, Word Ladders is purposefully designed to collect linguistic data from a broad spectrum of language users. The end-goal of Word Ladders is therefore that of collecting linguistic data from users (scientific objective) using a mobile app (gamified method). The application has been published on mobile stores in September 2023 and can be freely downloaded from Google Play at the following url: https://www.shorturl.at/kqAOS, and from App Store at the following url: https://www.shorturl.at/buIS2.

The game is based on the elicitation of word associations. Users are invited to construct ladders of words, by adding steps above and below an initial prompt. The rules of the game are hereby summarized (exactly as shown to the players):

1. The goal is to construct a ladder by adding words on each step. However, not all words will earn you points, so choose wisely!
2. Begin by adding increasingly more general words above the given prompt. Let’s say the starting word is “APPLE.” Above “APPLE,” you can add a word like “FRUIT.” Then, above “FRUIT,” you can add “FOOD.” Finally, above “FOOD,” you can add “OBJECT.” Remember, you’re thinking, “The word that I just put describes a type of …”
3. Each word you add should be more general than the previous one. In our example, APPLE is a type of FRUIT, which is a type of FOOD, which is a type of OBJECT.
4. On the steps below the initial word, you can be more specific and precise. For instance, you can add “GRANNY SMITH”, which is a specific type of apple, below “APPLE.”
5. The longer your ladder, the more points you earn! So, challenge yourself and your friends to build the longest ladder you can.
6. To help you come up with good words, think about the category or type that the word you just placed belongs to.
7. If you’re unsure about how to play, we have a graphic tutorial available for you to watch. It will guide you through the gameplay and give you a better understanding of the rules.

The Word Ladders platform offers three distinct game modes: individual, challenge, and group. In all three modes, the objective remains the same (**Figure 1**): building the longest ladder of valid words, starting from a given prompt. The time allocated to complete a ladder is uniformly set at 120 seconds across all modes.

In the individual mode, the user competes against the game system and therefore must construct the longest ladder within 60 seconds. This time duration has been changed to 120 seconds in the published version of the application, based on feedback provided by participants to the present case study.

The challenge mode enables users to challenge either a known user or a randomly chosen online user. This mode comprises three rounds, with each round constituting a match, where participants must complete a ladder. The user who creates the longest valid ladder on two out of 3 matches is the winner.

Lastly, the Group mode allows multiple users to engage in simultaneous gameplay, with all participants competing against each other. The user who constructs the longest valid ladder secures the game win.

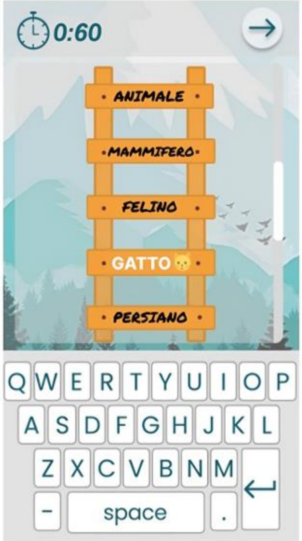

Figure 1. Screenshot of the visual instructions provided in the app Word Ladders, Italian version, to explain the game rules to players. The screenshot shows a match played on the prompt “GATTO” (cat). The player has added the following words above the prompt: “FELINO” (felix), “MAMMIFERO” (mammal) and “ANIMALE” (animal). The player has also added the following (more specific) word below the prompt: “PERSIANO” (Persian). (Source: Word Ladders, https://www.shorturl.at/kqAOS / https://www.shorturl.at/buIS2)

The ladder is evaluated against an external knowledge base: MultiWordNet 1.5.0, the multilingual lexical database aligned with Princeton WordNet 1.6. MultiWordnet is hosted by FBK Foundation in Trento, Italy and is licensed under a Creative Commons Attribution 3.0 Unported License. Further technical documentation is provided by Bolognesi et al. (2024).

To assess the quality of the ladders and assign ratings and scores to users, two parameters are employed: the number of words entered and the relevance of the entered words.

Word Ladders serves therefore as a gamified approach for gathering linguistic data, facilitating subsequent research and analyses on language use. Furthermore, it qualifies as a serious game due to its objective of not only integrating gamification elements into scientific tasks, but also striving to enhance and expand players’ active vocabulary.

Roadmap to Investigate the Advantage of Gamification through Word Ladders

The current investigation focuses on individuals with no known health conditions and centers specifically on a linguistically and cognitively complex task of word generation. In this task, participants are required to keep in mind explicit semantic rules to produce valid word associations. Given the task’s complex and repetitive nature across numerous trials, we anticipate that the implementation of gamification techniques will yield noteworthy advantages compared to a conventional online survey format.

Based on these observations, we can formulate a strong hypothesis and a moderate hypothesis, on the benefits of gamified approaches to the present linguistic data collection in relation to data quality. The strong hypothesis posits that the linguistic data hereby obtained through gamified techniques may demonstrate superior quality in comparison to data collected within a conventional experimental environment, specifically when involving healthy adult participants as research subjects. Conversely, the more moderate hypothesis suggests that the quality of linguistic data acquired through gamification has a comparable quality with that obtained through traditional survey-based methods. Given the task nature and the targeted research participants, who are university students accustomed to institutional settings, we are inclined towards adopting the moderate hypothesis for the present study.

Our second hypothesis centers around the user experience. As explained in the introduction, gamified approaches are perceived to offer a superior user experience compared to conventional survey forms commonly employed in institutional settings and academic crowdsourcing tasks. Gamification, as defined, involves incorporating elements of game design into non-game contexts. Hence, usability and gamification are closely intertwined concepts. By designing the interface to include elements that enhance engagement, attention levels, and usability, several benefits of gamification can be realized. Notably, it is worth considering that certain studies have demonstrated how the advantages of gamification extend beyond the immediate task completion timeframe, extending into long-term engagement. For instance, investigations such as "The Great Brain Experiment" have indicated that imposing a time constraint, even if brief (less than five minutes), heightens participants' inclination to engage repeatedly over an extended duration, as explained in a media release by UCL (UCL News, 2013). Based on these observations, we posit that participants utilizing a gamified technique to accomplish a linguistic task are likely to report a more favorable user experience in comparison to participants utilizing a traditional online survey.

Lastly, it is worth noting that gamification remains a subject of ongoing investigation. Despite numerous tests and research conducted in this domain, certain aspects still warrant further examination. One such aspect pertains to the role of competition between users, within gamified approaches, which has received limited scrutiny. We contend that competition between users serves as a crucial factor capable of motivating players to engage in assigned tasks, and this benefit can be quantified through reported motivation scores. Consequently, our third hypothesis posits that users will exhibit higher levels of self-motivation to participate in the task when playing in a challenge mode, against online adversaries, compared to users who play the online game in individual mode, and to the control group of users who completed the same task via a conventional web-based survey. This aspect carries significant importance as it suggests the potential scalability of data collection facilitated by gamified approaches in contrast to conventional methods. Put differently, if the incorporation of gamified elements, such as the competitive aspect that fosters direct comparisons among users, enhances user motivation, it can effectively promote an easily scalable mechanism for involving new users within the gaming platform.

METHODS

To test the impact of gamification on linguistic data collection we used a beta version of Word Ladders, directly distributed to the testers and not yet available on the Google Play Store. The same linguistic task involved in the game Word Ladders was also implemented in a classic survey, on the platform Qualtrics. Here, each trial consisted of an open-ended question in which participants were asked to construct a ladder starting from a given prompt. **Figure 2** displays a screenshot of the linguistic task implemented through Word Ladders (on the left) and a screenshot of the task implemented on the Qualtrics survey (on the right).

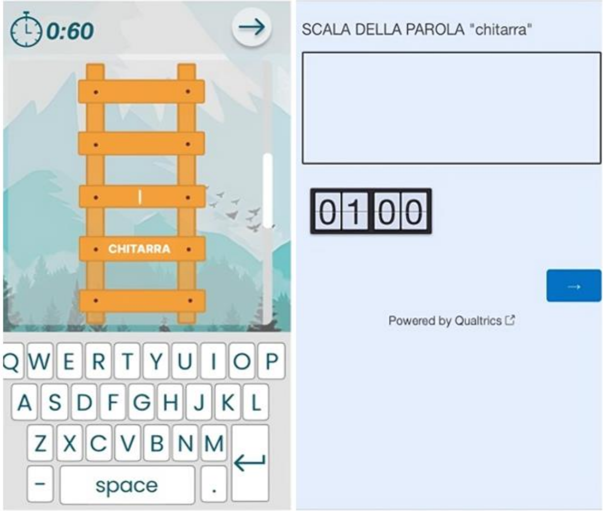

Figure 2. Screenshots of linguistic task to be done by experimental group & control group of participants (Source: Word Ladders, https://www.shorturl.at/kqAOS / https://www.shorturl.at/buIS2)

The detailed instructions provided to the experimental group of participants as well as to the control group, are reported in the online repository that contains also all the analyses hereby described, which were conducted on open-source platform for statistical analyses Statistics Kingdom (https://www.statskingdom.com/index.html) and validated with the software R (Core Team). Data here reported and interpreted is stored in Open Science Framework online repository: https://osf.io/26rh3/.

The current study encompassed two experimental phases:

1. An initial segment, wherein participants completed the linguistic task either on the Word Ladders app (experimental group) or the Qualtrics form (control group)
2. A subsequent segment, wherein all participants completed a survey containing questions regarding their evaluations of the performed task.

The experimental design employed to address our research questions encompasses two distinct measures: a direct measure of data quality, which evaluates the quality of words added by participants in response to initial prompts within the Word Ladders game or the Qualtrics form (for **H1**) and an indirect measure of overall experience, comprising participants' evaluations provided on a five-point Likert scale (for **H2** and **H3**).These judgments are gathered via a survey, also implemented on Qualtrics.

The quality of the data produced by the groups of participants (used to investigate **H1**) was measured based on two scores:

1. The length of the produced ladders, indicating the number of words entered in response to each given prompt.
2. The accuracy of the produced words, namely the number of relevant words entered for each ladder. In this context, relevance is determined by the presence of a hypernym or hyponym relationship with the prompt word, as established by the WordNet dataset (Fellbaum, 2005), which serves as our primary baseline for the evaluation of the data produced by app and form users.

The user experience and user motivation reported by the two groups of participants were measured based on a survey filled by both groups of participants, which was an adapted version of system usability scale (SUS) (Brooke, 1996). The minor modifications were necessary to tailor the survey to our specific research objectives (for instance, the survey was provided in Italian to participants). In the survey, participants were asked to rate their agreement on a scale from one to five, where one represented complete agreement and five represented complete disagreement.

The following questions were presented in the survey:

Q1. I am inclined to play this game again/perform this task again.

Q2. I found the game/form easy to use.

Q3. I believe most people would quickly learn to use this game/form.

Q4. Interacting with the platform did not raise any doubts while playing the game/performing the task.

Q5. I needed to learn many things before I could play the game/perform the task.

Q6. The game/task instructions were easy to comprehend.

Participants

The test was conducted with four classes of undergraduate, first year linguistics for computer science students, from the University of Bologna, in compliance with GDPR. A total of 189 students took part to the experiment. Students, male and female between 20 and 30 years old, were tested during four sessions, carried out between 12 December 2022 and 8 March 2023. Their participation in the experimental session was voluntary and free. Participants were invited to take part in the gaming session as well as in the appreciation survey (second phase of the experiment). However, given the voluntary nature of the testing, not all participants filled the appreciation survey.

Table 1 summarizes the participants involved in the phases of the experiment, divided by condition.

Table 1. Overview of participants involved in experimental sessions

Condition: Word Ladders app		Condition: Online form (control group)	
Individual mode: Single player	47 participants	Qualtrics survey, same prompts of Word Ladders app, & individual mode only	91 participants
Challenge mode: Two players on same prompts against each other	51 participants		
Total	98 participants		91 participants
Participants who filled appreciation survey after gaming session	73/98 participants		60/91 participants

Most students had Italian as their native language but in a few cases the user's reported native language was Albanian, Chinese, German, Japanese, Russian, Sinhalese, Spanish, and Ukrainian.

We did not operationalize this as a variable because all students were attending an Italian university course taught in Italian language. We therefore assume that their linguistic knowledge in Italian is fluent and may not constitute a problem for the task.

Materials

Words used as prompts for the task were 27 highly frequent Italian nouns, such as "chitarra" (guitar), "pavone" (peacock) or "doccia" (shower). Words were presented in randomized order to participants. The full list is provided in **Appendix A**, followed by its English translation. Words were presented in randomized order to participants.

Procedure

Participants were divided into three groups, which we called: the Individual app group, the challenge app and the form group (control). Participants took the test using their own cell phones, accessing either the QR code for downloading the Word Ladders apk, or by accessing the QR code to fill the task on Qualtrics (form group). The testing sessions had a total duration of about two hours each, which were necessary for explaining the rules (20 minutes), downloading the game or reaching the form (five minutes), filling the informed consent and data treatment policies (five minutes), playing the games or the task on the form, by building ladders on the word prompts listed above (60 minutes), and finally filling the survey on motivation and usability (20 minutes) and debriefing informally with the analysts, about the task (10 minutes). **Figure 3** displays an infographic that summarizes the experimental design of the present study.

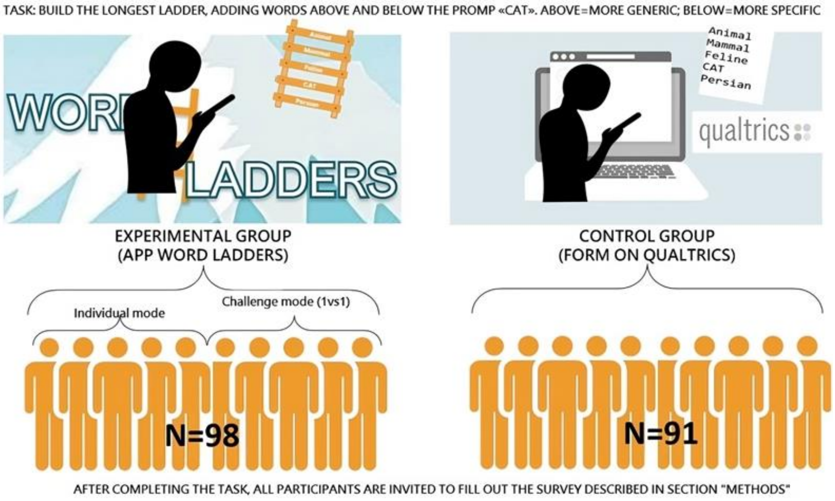
TASK: BUILD THE LONGEST LADDER, ADDING WORDS ABOVE AND BELOW THE PROMP «CAT». ABOVE=MORE GENERIC; BELOW=MORE SPECIFIC
WORD LADDERS
Animal
Mammal
Feline
CAT
Persian
qualtrics
EXPERIMENTAL GROUP (APP WORD LADDERS)
CONTROL GROUP (FORM ON QUALTRICS)
Individual mode
Challenge mode (1vs1)
N=98
N=91
AFTER COMPLETING THE TASK, ALL PARTICIPANTS ARE INVITED TO FILL OUT THE SURVEY DESCRIBED IN SECTION "METHODS"

Figure 3. Visualization of experimental design (Source: Authors)

ANALYSIS

Direct Measurement of Data Quality in Three Groups

As previously anticipated, to assess the quality of the word ladders (**H1**) produced by the two groups of participants (experimental group and control group) we considered two factors:

1. Average length of ladders produced by each user and
2. Average accuracy of terms used for each ladder produced by a user.

First, we analyzed the average length of the word ladders produced by participants during the experimental session and compared the means across the two main groups, namely app users and form users. Comparing app users (both, individual and challenge groups, n=98) against form users (n=91), the two sample Mann-Whitney U test shows that the app users altogether produced ladders that are on average significantly longer than those produced by form users (t=4.459, p<.001) with medium effect size (d=0.64). We used a Mann-Whitney U test because it does not assume normality distribution. As a matter of fact, a Shapiro-

Wilk test applied to the data distribution of the two groups revealed that neither of the two groups contains normally distributed data (p<.001 for both groups). The average length of ladders produced by app users is 5.448, while the average length of ladders produced by form users is 4.305. App users overall showed a higher engagement with the game, by producing, on average, longer word ladders.

Then we ran a more fine-grained analysis, comparing the three groups: Individual app, challenge app and form. The average length (number of words) of the ladders produced by the three groups are reported in **Table 2**. Overall, **Table 2** shows that in the challenge mode group, users entered more words compared to the individual mode group and to the control group of form users.

Table 2. Averages length of produced ladders for three groups

Group	Average ladder length
Individual app	4.895
Challenge app	5.957
Form	4.305

After checking the normality of the distribution of the three variables with a Shapiro-Wilk test and observing that the three distribution are normally distributed (p<.001), we ran a one-way ANOVA test of variance, which showed that the difference between the averages of some groups is big enough to be statistically significant (F=14.432, df=2,186, p<.001). The observed effect size *f* is large (0.39). The *η* equals 0.13. It means that the group explains 13.4% of the variance from the average. Specifically, a post-hoc Tukey HSD test showed the means of the following pairs are significantly different: individual app vs. challenge app, and challenge app vs. form. This pattern partially supports our **H1**: we obtained longer ladders, and therefore more data from app users compared to form users (control group). In particular, we obtained longer ladders from users who played in the challenge mode (interactive mode, against one another) compared to players who played individually or players who participated by filling the form. We found a general advantage of gamification, specifically when users play in interactive mode.

The second factor analyzed for investigating **H1** is the accuracy of the entered terms. By accuracy we mean whether the words entered are related to the prompt by a semantic relation considered valid, based on the game rules. For example, if the prompt word is 'cat' some relevant words are Persian, feline, mammal, animal, living being. Words like chair, whiskers, and dog are considered invalid. The validity of words is determined by their presence in the open-source database WordNet (Fellbaum, 2005).

Once again, we first compared the general mean accuracy of the ladders produced by app users (individual and challenge modes) to the mean accuracy of the ladders produced by form users. The average accuracy of app users (n=98) is 1.177, while the average accuracy of form users (n=91) is 1.370. The two sample Mann Whitney U test, used because of the non-normal distribution revealed by a Shapiro-Wilk test (p<.001 for both groups), shows that the form users altogether produced ladders that are on average significantly more accurate than those produced by app users (Z=-2.935, p=.020), although the effect is very small (Z=0.21) and the test barely approaches significance. This small observed difference between app users and form users should therefore be taken with caution.

We then proceeded to investigate the difference between the three groups. The average accuracy of the entered terms is shown in **Table 3**.

Table 3. Averages regarding accuracy of entered terms of three groups

Group	Mean score
Individual app	0.968
Challenge app	1.369
Form	1.370

The one-way ANOVA test of variance showed that the difference between the averages of some groups is big enough to be statistically significant (F=9.151, df=2,186, p<.001). The observed effect size *f* is medium (0.31). The *η* equals 0.09. It means that the group explains 9.0% of the variance from the average. Specifically, a post-hoc Tukey HSD test showed the means of the following pairs are significantly different: individual app vs. challenge app, and Individual app vs. form. This pattern partially supports our **H1**. In particular, it shows that the accuracy of the data produced by users in the challenge app group and in the form group is comparable between the two groups, while the accuracy of the data produced by the Individual app group is lower. We will discuss this pattern of results later.

Indirect Measures (Judgments) on Usability & Motivation

The questionnaire that participants filled out after completing the task on the app or on the Qualtrics form, was aimed at testing our second and third hypotheses, on users' perceived usability of the two tools for data collection (the app for groups individual app and challenge app, and the form for the control group) and their motivation to perform the task and repeat the experience.

As mentioned above, the survey was inspired by SUS. Users had to answer the questions by assigning a value from one to five, where one is totally agree and five is totally disagree. Overall, six questions asked in the survey tackled aspects of user experience, including users' motivation (notably, question 1).

First, to see whether the answers provided by participants presented strong correlations for some of the questions, we ran a correlation study, reported in **Table 4**, which displays Spearman's correlation coefficients within group, between questions.

Table 4. Spearman's correlation coefficients within group & between questions

	Q1	Q2	Q3	Q4	Q5	Q6
Individual app						
Q1	1.000					
Q2	0.576	1.000				
Q3	0.082	0.445	1.000			
Q4	0.154	0.506	0.170	1.000		
Q5	-0.352	-0.498	-0.294	-0.229	1.000	
Q6	0.239	0.443	0.375	0.412	-0.453	1.000
Challenge app						
Q1	1.000					
Q2	0.520	1.000				
Q3	0.227	0.591	1.000			
Q4	0.339	0.392	0.428	1.000		
Q5	-0.142	-0.409	-0.417	-0.316	1.000	
Q6	-0.142	-0.409	-0.417	-0.316	-0.316	1.000
Form group						
Q1	1.000					
Q2	0.265	1.000				
Q3	0.354	0.677	1.000			
Q4	0.032	0.168	0.273	1.000		
Q5	0.069	-0.165	-0.253	-0.316	1.000	
Q6	0.177	0.191	0.345	0.560	-0.197	1.000

Overall, there are no coefficients above 0.7, which may have suggested that two questions may have tackled the same issue, or that participants for some other reasons replied in the same way to different questions. There are however some positive and fairly strong correlations between questions. In the individual mode of the app, there was a significant correlation between question 1 and question 2. The correlation coefficient is: 0.576. That is, those who said they would play the game again also said the game was easy to play. In addition, a correlation was found between question 2 and question 4. The correlation coefficient is: 0.506. That is, those who found the game easy to play stated that the game platform did not cause them any problems or doubts while playing the game.

In the challenge mode of the app, there is as well a significant correlation between question 1 and question 2. The correlation coefficient is: 0.520. That is, those who stated that they would play the game again also stated that the game was easy to play. In addition, a correlation was found between question 2 and question 3. The correlation coefficient is: 0.591. That is, those who found the game easy to play stated that most people would learn to use this game very quickly.

In the form group there is a significant correlation between question 2 and question 3. The correlation coefficient is: 0.677. That is, those who found the task easy to perform stated that most people would learn to perform this task very quickly. In addition, a correlation was found between question 4 and question 6. The correlation coefficient is: 0.560. That is, those who said that the platform did not create doubts for them while performing the task also said that the game rules were clear and easy to understand.

We then proceeded to analyze and compare the average answers given to the questions by the three groups. In particular, we first tested for normality of the distributions, and based on the results we opted for a Mann Whitney U test (Wilcoxon rank-sum) to compare the overall mean values of app vs. form users, and then the Kruskal-Wallis H test, which does not assume normally distributed data, to compare the mean scores among the three groups: individual app, challenge app, and form group.

The means of the three groups are reported in **Table 5**, while the results of the statistical tests are reported in relation with each individual question.

Table 5. Mean scores given by participants within each group to each of questions in survey

Name	Data group					
	Q1	Q2	Q3	Q4	Q5	Q6
Individual app	2.943	2.314	1.914	2.257	3.286	1.857
Challenge app	2.421	1.921	1.684	2.632	3.921	1.892
Form group	2.368	2.912	2.316	2.228	3.544	1.702

Q1. "I am inclined to play this game again / perform this task again" (the lower the score, the higher the agreement with the statement).

A general comparison between app users and form users revealed no significant differences in relation to question 1, between the two main groups (Z=1.529, p=0.126).

However, a more fine-grained analysis in which the three groups are considered reveals interesting differences. Kruskal-Wallis H test indicated that there is a significant difference in the dependent variable between the different groups, χ(2)=6.98, p=.030. Post-hoc Dunn's test using a Bonferroni corrected alpha of 0.017 indicated that the mean rank of the pair individual app vs. form group is significantly different. The observed effect size *η*, however, is small, 0.039. Therefore, the group of participants who played the app in the Individual mode benefited of the gamification setting, compared to the control group. However, the group who played the in the challenge mode did not show a significant benefit over the control group of form users.

Q2. "I found the game/form easy to use" (the lower the score, the higher the agreement with the statement).

A general comparison between app users and form users revealed significant differences in relation to question 2, between the two main groups (Z=-5.095, p<.001). This shows that app users found significantly easier the interaction with the task, compared with form users.

Comparing the three groups, the Kruskal-Wallis test indicated that there is a significant difference in the dependent variable between the different groups, χ(2)=29.77, p<.001. Post-hoc Dunn's test using a Bonferroni corrected alpha of 0.017 indicated that the mean ranks of the following pairs are significantly different: individual app vs. form group, and challenge app vs. form group. In other words, both groups of app players (and in particular the group who used the app in the challenge mode) found the tool to be significantly easier to use, compared to the control group who performed the task on a Qualtrics form.

Q3. "I believe most people would quickly learn to use this game/form" (the lower the score, the higher the agreement with the statement).

Looking at the averages, we can say that the app is considered easier to use, in both its versions, than the form. Especially in the challenge version. Although the game screens, game rules, and game purpose are the same between the two modes, the challenge mode is perceived as easier.

Kruskal-Wallis test indicated that there is a significant difference in the dependent variable between the different groups, χ(2)=13.31, p=.001. Post-hoc Dunn's test using a Bonferroni corrected alpha of 0.017 indicated that the mean rank of the challenge app and the form group is significantly different.

In other words, the app users who used the challenge mode think that most people would easily learn to use the app, while this is not the case for the form group, who agree significantly less with the above statement compared to app users.

The last three questions of the liking test were focused on understanding whether even on a practical level gamification does not complicate the user experience and is not perceived as more tiring than traditional methods such as the form.

Q4. "Interacting with the platform did not raise any doubts while playing the game/performing the task" (the lower the score, the higher the agreement with the statement).

In this case, also the Kruskal-Wallis test indicated a non-significant difference in the dependent variable between the different groups. This means that the three groups gave on average similar responses to this question. In particular, the mean is below three, suggesting that there is more agreement than disagreement with the statement above.

Q5. "I needed to learn many things before I could play the game/perform the task" (the lower the score, the higher the agreement with the statement).

A general comparison between app users and form users revealed no significant differences in relation to question 1, between the two main groups (Z=0.623, p=0.533). However, Kruskal-Wallis H test indicated that there is a significant difference in the dependent variable between the different groups, χ(2)=6.93, p=.031. Post-hoc Dunn's test using a Bonferroni corrected alpha of 0.017 indicated that the mean rank of the pair individual app vs. challenge app is significantly different.

Interestingly, in this question the lower the score, the higher the agreement with the statement, but because the statements highlights difficulties found by users, the higher the mean, the lower the difficulties encountered. Therefore, app users in the challenge mode found significantly less difficulties, compared to the Individual mode users, with form users giving intermediate responses.

Q6. "The game/task instructions were easy to comprehend" (the lower the score, the higher the agreement with the statement).

A general comparison between app users and form users revealed no significant differences in relation to question 1, between the two main groups (Z=0.620, p=0.535).

The averages are not so different from each other. Given the low means across the three groups, it follows that the task instructions are equally easy to understand by all three groups.

DISCUSSION

Overall, we expected that data obtained through gamification techniques would exhibit (at least) the same quality as data collected through traditional methods, such as classic Qualtrics forms. The analyses conducted on the length and accuracy of the word ladders produced by app players vs. form users show an interesting scenario: app players, compared to the control group of form users tend to produce significantly more data through the gamified platform, but the quality of the data is slightly lower than the quality of the data produced by the form group. It should also be mentioned that the effect size of the latter finding, namely the higher accuracy of form users compared to app users, is very small, with the difference between the two groups appearing to be barely significant. In any case, we believe that this result may have been caused by the fact that the app version that was used to run the present study was a demo, not yet finalized and published on the mobile stores. Since then, revisions have been made to ensure that future data collections using this method will produce higher-quality data. Interestingly, the more fine-grained analysis in which we considered three groups of participants, namely app users in the mode Individual game, app users in the mode challenge game and form users (control group), we found that the data accuracy is the same between app users who played in the challenge mode, and form users. Conversely, app users who played in the Individual mode, even though they produced overall more data, the accuracy of the data they produced is significantly lower. This is interesting, because it partially support our moderate **H1**, namely the fact that a gamified approach can help collecting research data that are comparable in quality to the data collected in language sciences through classic online surveys, when the right modality of the game is implemented. In our case, the right modality proved to be challenge mode in which users could challenge one another to build the longest word ladders.

When comparing users' experience and motivation, the scenario is also quite complex. In particular, these two aspects of gamification were explored by our study by comparing overall the app users and the form users, on the same linguistic task, but also comparing three groups of users in a more fine-grained manner that allowed us to disentangle potential differences between different app modalities. In particular, we compared app players that performed the task in a modality that involved one-to-one challenges, and app players that performed the task in an individual modality. We ran this more fine-grained analysis because we expected some differences between the two sub-groups of app players and a stronger motivation due to social interaction and competition. Indeed, social mobile gaming proved to be a successful tool to motivate and engage users (Elciyar, 2015) but there is not specific study focused on social linguistics mobile games.

Surprisingly enough, our results show that app players, and therefore a gamified approach to the linguistic task to be performed, are not always more inclined to play again (increase in motivation, tackled in question 1) compared to the control group. Only the group of participants who played the app in the Individual mode benefited of the gamification setting, compared to the control group. However, the group who played the in the challenge mode did not show a significant benefit over the control group of form users. This finding is in line with empirical reviews on gamification showing that the increase in motivation is task-dependent (Lumsden et al., 2016) and that the game characteristics may be beneficial or deliver similar results as to more classic online surveys.

The usability of the gamified platform, the Word Ladders game, proved to be positive, and significantly higher than the usability of the more traditional Qualtrics online survey, for the linguistic task hereby investigated. Question 2 and question 3 showed that on average app users find the platform easy and easy to learn, compared to form users. In any case, neither of the platforms raised overall substantial doubts (question 4). Interestingly, however, app users in the challenge mode found significantly less difficulties, compared to the Individual mode users, with form users giving intermediate responses. This again suggests that an interactive game modality may stimulate users to be more motivated and positive toward the experience, compared to an individual game in which the game elements do not include this interaction between players.

CONCLUSIONS

Through our research, we aimed to examine a number of hypotheses about the benefits of gamification techniques in data collection for linguistic purposes, including their impact on data quality, user experience, and sustained motivation for continued engagement.

We were able to verify that competition is a gamification element that exerts a substantial influence on user experience. This impact is evident as users exhibit a greater inclination to engage in the challenge mode as opposed to the individual mode. Furthermore, the utilization of competition yields advantages in terms of ladder length, as users demonstrate heightened motivation to construct longer word ladders. Moreover, these benefits extend to usability, as the challenge mode is perceived as facilitating a more effortless gameplay experience overall.

Our efforts are far from concluded. The availability of the app on the major app stores ensures an increasing number of new users and a continuous flow of user data to evolve and integrate with those collected during our formal trials with a controlled set of participants. We are planning several further steps to widen the scope and strengthen the results obtained from our users. First of all, we plan to widen the variety of prompts, increase the number of game levels and work on the precision of the evaluations. Then, as we turn from a selected group of mostly ethnographically uniform university students to a more varied collection of members of the general population, we plan to increase the analysis and categorization of our result data according to a much ampler variety of users, especially regarding age, provenance and schooling level. Other types of variables in participants characteristics, including comfort with the technology, love for word games and richness of vocabulary, can then be tested against, too.

In fact, we believe we have studied a repeatable approach for collecting large quantities of *high-quality* data in the *linguistic domain*, leveraging on users’ engagement in participating to an interesting game and challenging each other. The combination of these two elements can be easily repeated to verify other research hypotheses in the same domain, as well as to collect data in totally different areas like education, psychology and hard sciences.

Author contributions: FG, MMB, ADI, & FV: design of experimental study & data collection; **FG & MMB:** data analyses; **FG:** drafted introduction, methods & analyses; **MMB:** wrote theoretical background, discussion, & conclusions and edited the introduction, methods and analyses; & **ADI & FV:** edited and revised the paper. All authors approved the final version of the article.

Funding: This article was funded by the European Union (GRANT AGREEMENT: ERC-2021-STG-101039777 awarded to Marianna Bolognesi). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgements: The authors would like to thank all participants who took part to the experimental sessions at the Departments of Modern Languages and Computer Sciences of the University of Bologna, Italy.

Ethics declaration: The authors declared that the study was approved by the Ethics Committee at University of Bologna on 20 May 2022 with approval number: 0121435. Written informed consents were obtained from the participants.

Declaration of interest: The authors declare no competing interest.

Data availability: Data are stored in the following online repository, on Open Science Framework: https://osf.io/26rh3/.

REFERENCES

AlSaad, F. M., & Durugbo, C. M. (2021). Gamification-as-innovation: A review. *International Journal of Innovation and Technology Management, 18*(05), 2130002. https://doi.org/10.1142/S0219877021300020

Benini, S., & Thomas, M. (2021). A critical review of research on gamification and second language acquisition. In M. Peteson, M. Thomas, & K. Yamazaki (Eds.), *Digital games and language learning* (pp. 9-46). Bloomsbury. https://doi.org/10.5040/9781350133037.ch-002

Bolognesi, M., Collacciani, C., Ferrari, A., Genovese, F., Lamarra, T., Loia, A., Rambelli, G., Ravelli, A. A., & Villani, C. (2024). *Word Ladders: A mobile application for semantic data collection*. arXiv:2404.00184 [cs.CL]

Bonetti, F., & Tonelli, S. (2021). Measuring orthogonal mechanics in linguistic annotation games. *Human-Computer Interaction, 5*, 265. https://doi.org/10.1145/3474692

Brooke, J. (1996). SUS–A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189-194). Taylor & Francis.

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. https://doi.org/10.1080/09588221.2019.1648298

DeRight, J., & Jorgensen, R. (2015). I just want my research credit: frequency of suboptimal effort in a non-clinical healthy undergraduate sample. *Clinical Neuropsychology, 29*(1), 101-117.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9-15). https://doi.org/10.1145/2181037.2181040

Elciyar, K. (2015). Social mobile gaming and user practices. *Online Journal of Communication and Media Technologies, 5*, 141-156. https://doi.org/10.30935/ojcmt/5680

Eryigit, G., Bektas, F., Ali, U., & Dereli, B. (2021). Gamification of complex morphology learning: The case of Turkish. *Computer Assisted Language Learning, 36*(8), 1421-1449. https://doi.org/10.1080/09588221.2021.1996396

Eryigit, G., Sentas, A., & Monti, J. (2023). Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering, 29*(4), 909-941. https://doi.org/10.1017/S1351324921000401

Fellbaum, C. (2005). WordNet and wordnets. In K. Brown, & R. E. Asher (Eds.), *Encyclopedia of language and linguistics* (pp. 665-670). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00946-9

Flatla, D. R., Gutwin, C., Nacke, L. E., Bateman, S., & Mandryk, R. L. (2011). Calibration games: Making calibration tasks enjoyable by adding motivating game elements. In *Proceedings of the ACM Symposium on User Interface Software and Technology* (pp. 403-412). ACM. https://doi.org/10.1145/2047196.2047248

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*(5), 847-857. https://doi.org/10.3758/s13423-012-0296-9

Gundry, D., & Deterding, S. (2019). Validity threats in quantitative data collection with games: A narrative survey. *Simulation & Gaming, 50*(3), 302-328. https://doi.org/10.1177/1046878118805515

Hammady, R., & Arnab, S. (2022). Serious gaming for behaviour change: A systematic review. *Information, 13*(3), 142. https://doi.org/10.3390/info13030142

Hawamdeh, M., & Soykan, E. (2021). Systematic analysis of effectiveness of using mobile technologies (MT) in teaching and learning foreign language. *Online Journal of Communication and Media Technologies, 11*(4), e202124. https://doi.org/10.30935/ojcmt/11256

Ishaq, K., Zin, N. A. M., Rosdi, F., Jehanghir, M., Ishaq, S., & Abid, A. (2021). Mobile-assisted and gamification-based language learning: A systematic literature review. *PeerJ Computer Science, 7*, e496. https://doi.org/10.7717/peerj-cs.496

Kim, Y., Kogan, V. V., & Zhang, C. (2023). Collecting big data through citizen science: Gamification and game-based approaches to data collection in applied linguistics. *Applied Linguistics*. https://doi.org/10.1093/applin/amad039

Krisbiantoro, B. (2020). The effectiveness of gamification to enhance students' mastery on tenses viewed from students' creativity. *Journal of Advanced Multidisciplinary Research, 1*(2), 73-97. https://doi.org/10.30659/jamr.1.2.73-97

Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G., & Mehr, S. A. (2023). How games can make behavioral science better. *Nature, 613*(7944), 433-436. https://doi.org/10.1038/d41586-023-00065-6

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games, 4*(2), e11. https://doi.org/10.2196/games.5888

McPherson, J., & Burns, N. R. (2007). Gs invaders: Assessing a computer game-like test of processing speed. *Behavior Research Methods, 39*(4), 876-883. https://doi.org/10.3758/BF03192972

Ogawa, H., Nishikawa, H., Tokunaga, T., & Yokono, H. (2020). Gamification platform for collecting task-oriented dialogue data. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 7084-7093).

Purgina, M., Mozgovoy, M., & Blake, J. (2020). WordBricks: Mobile technology and visual grammar formalism for gamification of natural language grammar acquisition. *Journal of Educational Computing Research, 58*(1), 126-159. https://doi.org/10.1177/0735633119833010

Qian, M., & Clark, K. R. (2016). Game-based Learning and 21st century skills: A review of recent research. *Computers in Human Behavior, 63*, 50-58. https://doi.org/10.1016/j.chb.2016.05.023

Reeves, B., & Read, J. L. (2009). *Total engagement: Using games and virtual worlds to change the way people work and businesses compete*. Harvard Business School Press.

Reiners, T., Wood, L. C., Gregory, S., & Teras, H. (2015). Gamification design elements in business education simulations. In M. Khoshrow-Pour (Ed.), *Encyclopedia of information science and technology*. IGI Global. https://doi.org/10.4018/978-1-4666-5888-2.ch298

Ritterfeld, U., Cody, M., & Vorderer, P. (2009). *Serious games: Mechanisms and effects*. Routledge. https://doi.org/10.4324/9780203891650

Ro, M., Brauer, M., Kuntz, K., Shukla, R., & Bensch, I. (2017). Making cool choices for sustainability: Testing the effectiveness of a game-based approach to promoting pro-environmental behaviors. *Journal of Environmental Psychology, 53*, 20-30. https://doi.org/10.1016/j.jenvp.2017.06.007

Sailer, M., Hense, J., Mandl, H., & Klevers, M. (2013). Psychological perspectives on motivation through gamification. *Interaction Design and Architecture(s) Journal, 19*, 28-37. https://doi.org/10.55612/s-5002-019-002

Sevastjanova, R., Jentner, W., Sperrle, F., Kehlbeck, R., Bernard, J., & El-Assady M. (2021). QuestionComb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling. *ACM Transaction on Interactive Intelligent Systems, 11*(3-4), 19. https://doi.org/10.1145/3429448

Sever, N. S., Sever, G. N., & Kuhzady, S. (2015). The evaluation of potentials of gamification in tourism marketing communication. *International Journal of Academic Research in Business and Social Sciences, 5*(10), 188-202. https://doi.org/10.6007/IJARBSS/v5-i10/1867

Short, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie B. (2023). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning, 36*(3), 517-554. https://doi.org/10.1080/09588221.2021.1933540

Smiderle, R., Rigo, S. J., Marques, L. B., de Miranda Coelho, J. A. P., & Jaques, P. A. (2020). The impact of gamification on students' learning, engagement and behavior based on their personality traits. *Smart Learning Environments, 7*, 3. https://doi.org/10.1186/s40561-019-0098-x

Smith, A., Legaki, N., & Hamari, J. (2022). Games and gamification in flipped classrooms: A systematic review. In M. Bujic, J. Koivisto, & J. Hamari (Eds.), *Proceedings. 6th International GamiFIN Conference 2022 (GamiFIN 2022)* (CEUR Workshop Proceedings, Vol. 3147, pp. 33-43). CEUR-WS.org.

UCL News. (2013). *The Great Brain experiment: Crowdsourcing data on how we think and act*. https://www.ucl.ac.uk/news/2013/mar/great-brain-experiment-crowdsourcing-data-how-we-think-and-act

Verduin, M. L., LaRowe, S. D., Myrick, H., Cannon-Bowers, J., & Bowers, C. (2013). Computer simulation games as an adjunct for treatment in male veterans with alcohol use disorder. *Journal of Substance Abuse Treatment, 44*, 316-322. https://doi.org/10.1016/j.jsat.2012.08.006

Waluyo, B., Phanrangsee, S., & Whanchit, W. (2023). Gamified grammar learning in online English courses in Thai higher education. *Online Journal of Communication and Media Technologies, 13*(4), e202354. https://doi.org/10.30935/ojcmt/13752

Welbers, K., Konijn, E., Burgers, C., & de Vaate, A. (2019). Gamification as a tool for engaging student learning: A field experiment with a gamified app. *E-Learning and Digital Media, 16*(2), 92-109. https://doi.org/10.1177/2042753018818342

Wright, W., & Bogost, I. (2007). *Persuasive games: The expressive power of videogames*. MIT Press.

Zheltukhina, M. R., Kislitsyna, N. N., Panov, E. G., Atabekova, A., Shoustikova, T., & Kryukova, N. I. (2023). Language learning and technology: A conceptual analysis of the role assigned to technology. *Online Journal of Communication and Media Technologies, 13*(1), e202303. https://doi.org/10.30935/ojcmt/12785

Zichermann, G., & Cunningham, C. (2011). *Gamification by design: Implementing game mechanics in web and mobile apps*. O'Reilly Media.

APPENDIX A

List of Stimuli

1. Arte (art)
2. Bambino/bambina (baby)
3. Banana (banana)
4. Bocca (mouth)
5. Camicia/maglia (shirt)
6. Chitarra (guitar)
7. Compito (task)
8. Contentezza (contentment)
9. Doccia (shower)
10. Festa (party)
11. Giallo (yellow)
12. Giornale (newspaper)
13. Hotel (hotel)
14. Margherita (daisy)
15. Medico/dottore (doctor)
16. Mela (apple)
17. Miele (honey)
18. Orologio (clock)
19. Pane (bread)
20. Pavone (peacock)
21. Pizza (pizza)
22. Specchio (mirror)
23. Statua (statue)
24. Tavolo (table)
25. Tazza (cup)
26. Telefono (phone)
27. Volpe (fox)