Bräuer, Paula; Mazarakis, Athanasios

Article — Published Version

How to Design Audio-Gamification for Language Learning with Amazon Alexa? — A Long-Term Field Experiment

International Journal of Human–Computer Interaction

Suggested Citation: Bräuer, Paula; Mazarakis, Athanasios (2024) : How to Design Audio-Gamification for Language Learning with Amazon Alexa? — A Long-Term Field Experiment, International Journal of Human–Computer Interaction, ISSN 1532-7590, Taylor & Francis, London, Vol. 40, Iss. 9, pp. 2343-2360,
https://doi.org/10.1080/10447318.2022.2160228

This Version is available at:
http://hdl.handle.net/11108/556

International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

How to Design Audio-Gamification for Language Learning with Amazon Alexa?—A Long-Term Field Experiment

Paula Bräuer & Athanasios Mazarakis

To cite this article: Paula Bräuer & Athanasios Mazarakis (2024) How to Design Audio-Gamification for Language Learning with Amazon Alexa?—A Long-Term Field Experiment, International Journal of Human–Computer Interaction, 40:9, 2343-2360, DOI: 10.1080/10447318.2022.2160228

To link to this article: https://doi.org/10.1080/10447318.2022.2160228

Published online: 29 Dec 2022.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2024, VOL. 40, NO. 9, 2343–2360
https://doi.org/10.1080/10447318.2022.2160228

How to Design Audio-Gamification for Language Learning with Amazon Alexa?—A Long-Term Field Experiment

Paula Bräuer[a] and Athanasios Mazarakis[b]

[a]Computer Science Department, Kiel University, Kiel, Germany; [b]ZBW-Leibniz Information Centre for Economics, Kiel, Germany

ABSTRACT

Gamification can increase motivation in learning, and intelligent virtual assistants (IVAs) can support foreign language learning at home. However, there is a lack of design concepts to motivate learners to practice with their IVA. This study combines both concepts and analyzes if audio-gamification can increase engagement to address this research gap. To this end, a one-year long-term field experiment with 230 subjects using a German language learning skill for Amazon Alexa was conducted. A between-subjects design determined differences in learning behavior and learning outcomes between a control group and two gamified groups (achievements and leaderboard). The findings reveal a positive effect on the number of translated vocabulary and learning success. However, only in the group with a leaderboard was a statistically significant effect on the number of translated vocabulary found. These findings imply that audio-gamification can be a helpful tool for increasing motivation to use IVAs for foreign language learning.

1. Introduction

Individuals of many ages and backgrounds are interested in learning another language or may be obliged to do so (Young, 2014). However, learning a language takes time and effort, and language students often do not receive enough exposure or practice in the language they want to learn (Govender & Arnedo-Moreno, 2021). The development of new interaction methods with technological systems, such as intelligent virtual assistants (IVAs), can help learners with these challenges (Dizon, 2021; Istrate, 2019; Skidmore & Moore, 2019). In this respect, the term IVA is used interchangeably in the literature with terminologies, such as intelligent personal assistant (IPA), conversational agent (CA), conversational user interface (CUI), virtual personal assistant (VPA), or voice-enabled assistant (VA), to mention a few (Cowan et al., 2017). IVAs can be defined as "interfaces that enable users to interact with smart devices using spoken language in a natural way and provide a singular response to a query similar to speaking to a person" (McTear et al., 2016; Nobles et al., 2020).

IVAs like Amazon Alexa, Apple Siri, or Google Assistant are now available on various devices, such as smartphones and speakers. As a result, the use of speech to interact with automated systems has grown in popularity (Clark et al., 2019; Lopatovska et al., 2020). But despite their increasing prevalence, most individuals do not use IVAs regularly because they miss integration across devices and the ability to customize according to their regular usage habits (Cowan et al., 2017). However, regular practice and repetition are essential to learn a new language. Therefore, for students to learn successfully with an IVA, student acceptance and satisfaction with the IVA are critical (Babic et al., 2018).

Moreover, IVAs are already used for learning new languages (Benner et al., 2022; Dizon, 2021; Istrate, 2019; Skidmore & Moore, 2019). By interacting with an IVA, for example, the listening comprehension, as well as the pronunciation, can be improved. Another advantage is that IVAs offer an easy way to train a new language interactively outside the usual classroom scenario (Istrate, 2019).

Even though the topic of IVA is increasing its relevance for research, there is currently a significant research gap in terms of design principles for IVAs (Clark et al., 2019). Qualitative studies, for example, show that many users rate IVA feedback often as inadequate and that feedback needs hints for improvement in a transparent way to achieve meaningful feedback (Lopatovska et al., 2020; Luger & Sellen, 2016). Thus, designing IVAs to encourage users to interact with them over time is challenging.

One approach that has been successfully applied in various contexts in Human–Computer Interaction to increase user motivation is gamification (Hamari et al., 2014; Koivisto & Hamari, 2019). According to Deterding et al. (2011), gamification is defined as "the use of game design elements in nongame contexts." Besides the potential to increase user motivation, gamification has also been established as a helpful tool in the field of education (Bai et al., 2020; Dichev & Dicheva, 2017; Dicheva et al., 2015; Majuri et al., 2018). Aside from the most commonly observed

CONTACT Athanasios Mazarakis a.mazarakis@zbw.eu ZBW-Leibniz Information Centre for Economics, Düsternbrooker Weg 120, Kiel, 24105, Germany

effects on motivation and performance, an increase in learning success has been repeatedly demonstrated in education with the help of gamification (Denny et al., 2018; Ortiz-Rojas et al., 2019).

Visual elements are common when designing and implementing a gamification approach; badges or leaderboards are typically used in a classroom setting (Christy & Fox, 2014; Sailer & Sailer, 2021; Tan & Hew, 2016). However, visual components should be avoided when integrating such game design elements into applications for an IVA since many end devices do not provide a screen, like Alexa Echo Dot or Google Nest Mini (Kinsella, 2019).

When designing dialogs, it is best to keep things as simple as possible in accordance with current design guidelines (Murad et al., 2018). However, integrating purely acoustically presented game design elements is a challenge because, due to fundamental differences in vision and hearing, elements from classic visual computer games cannot be transferred without considerable alterations to their properties (Friberg & Gärdenfors, 2004). Audio games, which, unlike traditional video games, do not have visual elements, are an example of how games can be created without a visual component (Garcia et al., 2013). Initial results have already shown that gamification can also be implemented purely acoustically without visual elements (Bräuer & Mazarakis, 2022), thus creating audio-gamification (Mazarakis, 2021). In a laboratory setting, Bräuer and Mazarakis found that the speed of processing household tasks was increased compared to a control group by audio-gamification.

As a research question, we aim to find support with this study that by combining the potentials of IVAs and audio-gamification in the field of language learning, positive effects on motivation and learning success can be achieved. In particular, unlike in previous studies (Bräuer & Mazarakis, 2022; Dizon, 2020; Moussalli & Cardoso, 2020; Sailer et al., 2017), the aim is not to work with students in a laboratory setting but to collect real-world data with a long-term field experiment that reflects real-world usage behavior. This article aims to contribute to this goal by investigating the integration of audio-implemented game design elements with IVAs. This approach is central because gamification studies usually rely on relatively short-term studies (Cermak-Sassenrath, 2019), and studies about learning with IVAs are typically conducted in lab settings (Dizon, 2020; Hsu et al., 2021; Pyae & Scifleet, 2018; Wu et al., 2020). By developing a dedicated application for Amazon Alexa, a so-called Alexa skill, it was possible to record usage behavior in detail in the various conditions. The evaluation of this quantitative data provides new insights that previous studies could not capture through surveys and self-reporting by users.

The rest of the article is organized as follows: The next chapter provides a brief theoretical background for the research. Then, the results of an empirical study are presented in the third chapter and discussed in the following. Next, limitations and future research approaches are referred. Finally, the results of the article are briefly summarized.

2. Background

IVAs in the context of language learning offer much potential for research (Dizon, 2021). At the same time, gamification in education is a popular topic in human–computer interaction (HCI) research (Dicheva et al., 2015; Seaborn & Fels, 2015). This section will review the related work in two aspects: language learning with virtual assistants and language learning in combination with gamification.

2.1. Learning foreign languages with intelligent virtual assistants

Many low-cost self-study tools exist to learn a foreign language, regardless of whether it is about reading or writing, but opportunities to practice speaking skills are far more limited (Ruan et al., 2021). IVAs can currently speak a variety of languages and can therefore be used as a language-learning tool. For example, Amazon Alexa is multilingual in nine languages (*Develop Skills in Multiple Languages | Alexa Skills Kit*, 2021), Google Assistant in 16 languages (*Change the language of Google Assistant—Android—Google Nest Help*, 2021), Siri is able to communicate in 21 different languages (*IOS and IPadOS-Feature Availability*, 2021), and the number of languages spoken by the assistants is constantly increasing. According to de Barcelos Silva et al.'s (2020) systematic literature review, education, in particular, could be a potentially relevant application area for IVAs, but it has not yet been sufficiently investigated.

Teaching with the help of IVAs poses many advantages. Speech recognition offers the possibility to train and specify the pronunciation of a newly learned language. Regular exchange in a foreign language with the IVA can also make learners aware of gaps in their linguistic knowledge (Dizon, 2020). In addition, meeting with a teacher to perform exercises is no longer necessary, but the student can train at home alone with the voice assistant (Istrate, 2019). So learning autonomy could be aided by IVAs, which provide learners with opportunities for language practice in a low-stress setting (Dizon, 2021). Also, most first-timers shy away from speaking in front of others when learning a new language (Namaghi et al., 2015). IVAs, on the other hand, could provide a friendly, non-intimidating environment for spoken language practice. In addition, the IVA's continuous availability also allows for responding to the learner's personal needs. The learner is not required to follow a predetermined lesson plan at a predetermined time but can practice with the IVA at any time of the day.

The scientific literature on IVAs in the context of language learning has focused on learners' attitudes and experiences with learning technologies. Several studies have reported difficulties for second language speakers (L2) related to being understood by an IVA (Chen et al., 2020; Dizon, 2017; Dizon et al., 2022; Pyae & Scifleet, 2019). Moussalli and Cardoso (2020) deal with this problem in more detail. Their results show that L2 learners have in contrast no problems understanding Alexa and that it adapts well to their accented speech. The results also show that L2 learners use a variety of strategies to address the

communication difficulties they experience with Alexa, e.g., repeating or rephrasing when problems in communication occur.

IVA interaction is often task-oriented and takes the form of questions and responses (Cowan et al., 2017). Vocabulary learning is an example of where this principle can be put to good use. Skidmore and Moore (2019) developed an Alexa skill called "Japanese Flashcards." The authors consider areas of the Alexa development process that limit the facilitation of language learning, particularly the lack of multilingual speech recognition, and offer solutions to these limitations. The flashcard system applied in their article is commonly used for vocabulary learning. The concept is that the learner is presented with one side of the flashcard and tries to remember the content of the other side. Despite the fact that the Alexa skill has yet to be evaluated, the authors believe Alexa can contribute to the developing field of "voice-assisted language learning" by providing valuable learning paradigms, such as conversational role-play and pronunciation training (Skidmore & Moore, 2019).

In two similar studies, Ruan et al. (2019) compared a text-based chatbot app to a flashcard app. Even though both systems used the same algorithm to sequence the material, students recognized (and remembered) more correct answers when using the chatbot than when using the flashcard app. While using a chatbot was more time-consuming, in a second study, students spent 2.6 times more time on their own initiative using a chatbot rather than flashcards, indicating a strong preference for occasional learning. Also, the results of the second study revealed that chatbots outperformed flashcards in terms of learning gains in recall (Ruan et al., 2019).

Dizon (2020) investigated the effects of using Alexa on L2 English students in a quasi-experiment with an experimental group that received a 10-week treatment and a control group that did not receive such a treatment. The emphasis was on improving listening comprehension and speaking proficiency. According to the author, the experimental group was able to make more significant gains in L2 speaking proficiency. However, no significant difference was found when comparing improvements in L2 listening comprehension. The results of Dizon (2020) get support from Hsu et al. (2021), whose study also only showed an improvement in pronunciation but not in listening comprehension. Both studies used a setting where subjects performed various tasks with an IVA to practice a new language. For this purpose, the system language of the IVA was set to the foreign language to be learned. In addition, the subjects were given various instructions on how to interact with the device.

An important issue in working with IVAs for foreign language learning is the difference in communication behavior between native (L1) and L2 speakers. One finding by Wu et al. (2020) was that L2 users were seeking to lower their amounts of language production in conversation with the IVA. This was accompanied by frustration, having to reformulate queries from scratch. In contrast, L1 speakers tend to pay more attention to the conciseness of their phrases to counteract the system limitations of the IVAs.

Pyae and Scifleet (2019) found similar results in their study, in which L2 speakers reported being able to easily learn to operate the IVA they were using but encountered considerable difficulty in obtaining the necessary words and formulating instructions in a way that the IVA understood. In addition, Wu et al. (2020) showed that L2 speakers favored smartphones because the visual feedback allowed for the diagnosis of communication problems while offering time to analyze the results. On the other hand, L1 speakers, who thought audio feedback was sufficient, chose smart speakers. Pyae and Scifleet (2018) showed in a study that the differences between L1 and L2 could also affect user experiences. In comparison, L1 speakers found IVAs easier and simpler to use and had more confidence in using them.

In a comprehensive literature review on IVAs for foreign language learning, Dizon (2021) identified opportunities and challenges in this area. A key finding of the author's literature review was that many subjects in qualitative surveys complained that feedback in language learning through IVAs was very limited. Furthermore, there are instances where language is studied solely for testing purposes rather than for its communicative function or cultural richness. When students learn a language for its utility rather than its fun, they face obstacles, such as a lack of interest and motivation, which leads to a lack of practice (Govender & Arnedo-Moreno, 2021). By using various game design elements, gamification could assist in overcoming the problem of low motivation and generate new opportunities to provide users with relevant feedback (Mazarakis, 2015, 2021). To address this research and design gap, our study examines audio-gamification in the context of IVAs for foreign language learning.

2.2. Gamification in foreign language learning

Gamification has been defined as using game design elements in nongame contexts (Deterding et al., 2011). In learning, gamification is a design process in which game design elements are added to modify existing learning processes (Sailer & Homner, 2020). As a result, gamification became of significant interest to educators who have explored its potential to improve student learning (Bai et al., 2020; Dichev & Dicheva, 2017; Dicheva et al., 2015; Legaki et al., 2020; Majuri et al., 2018). In practice, there are also several examples of learning applications that use gamification. The best-known example is probably Duolingo (Adams, 2019), an application that can be used to learn a wide variety of foreign languages.

According to the theory of gamified learning, gamification influences learning outcomes by improving relevant attitudes and behaviors (Sailer & Homner, 2020). In addition, gamification is typically used to improve user motivation for a specific task or activity (Hamari et al., 2014; Sailer et al., 2017).

Motivational psychology theories and concepts have been particularly popular among HCI scientists for describing

and analyzing games and understanding what constitutes engaging player-computer interaction (Tyack & Mekler, 2020). In gamification research, Deci and Ryan's self-determination theory (Ryan & Deci, 2000) is commonly used to explain motivation (Seaborn & Fels, 2015). Motivation can be either extrinsic or intrinsic, according to this concept. Intrinsic motivation is characterized by doing something because it is inherently interesting or fun. On the other hand, extrinsic motivation refers to doing something because it leads to a definable result (Ryan & Deci, 2000, 2017). The theory of basic psychological needs is a derivative theory of self-determination theory (Deci & Ryan, 2000). According to this theory, there are three basic needs: experience of competence, a need for challenge, and feelings of effectance (Deci & Ryan, 2000); autonomy, a sense of volition or willingness when doing a task (Deci & Ryan, 2000); and social relatedness, which is experienced when a person feels connected with others (La Guardia et al., 2000). The satisfaction of these three needs might lead to intrinsic motivation. The concept argues that an individual's environment has an impact on their satisfaction with basic needs. This influence can be both positive and negative, contributing to or limiting the satisfaction of needs (Vansteenkiste & Ryan, 2013). However, it can be assumed that a significant part of the motivation in daily tasks is extrinsic and that gamification is externally endorsed (Mazarakis & Bräuer, 2022; Mekler et al., 2017). Nevertheless, it is not impossible that in some cases, an extrinsic motivation could be converted into an intrinsic motivator.

A literature review by Majuri et al. (2018) shows that gamification in the educational context achieved similar results as in other contexts. Most of the studies examined psychological aspects of the effect of gamification and mainly reported positive results. In addition, most studies found a positive effect on measurable educational outcomes. There are a variety of metrics to measure such an effect (Looyestyn et al., 2017), including, but not limited to course grades (Denny et al., 2018; Dichev & Dicheva, 2017), time spent (Landers & Landers, 2014; Monterrat et al., 2015), or volume of contributions (Bouchrika et al., 2021; Tejedor-Garcia et al., 2020). Since successful gamification involves repetition of the desired results (Robson et al., 2015), it makes perfect sense to apply it to vocabulary learning, where repetition is also essential to improve recall.

Dehghanzadeh et al. (2021) conducted a literature review on gamification in the context of learning English as a foreign language, which included 22 publications. Besides grammar or pronunciation training, vocabulary learning was by far the most common application in which gamification was used in the studies considered. None of the publications considered there could show negative results from the integration of gamification. Slightly more than half of the studies stated that the learners' experience was positively influenced by the integration of the game design elements. In terms of the literature reviewed, the authors criticize, among other things, the fact that many of the studies were conducted without a proper control group and that the data was mostly collected in a very short period of time, which can be considered as important research gaps.

How gamification can be used to enhance IVA-supported learning of English as a foreign language was investigated in a study by Tejedor-Garcia et al. (2020). The authors report the development and evaluation of a gamified application for practicing English vocabulary pronunciation. For voice recognition, an app was created for smartphones that interact with Google Assistant. To encourage users to stay motivated, challenges were developed in which they could compete. A comparison was made with data from an earlier study (Tejedor-Garcia et al., 2016) to see how the game design elements affected motivation and performance in pronunciation practice. Learning success and motivation exhibited positive trends in the comparison. Visual representation was employed to implement the game design elements (challenge, leaderboard, points, avatar, and badges). The results suggest that gamification is helpful in this context. However, it remains an open question whether gamification would have a positive effect without visual components.

So far, it is unclear which game design elements should be used, both for IVAs (Bräuer & Mazarakis, 2022) and in general (Mazarakis & Bräuer, 2022; Mekler et al., 2017). A general analysis of game design elements used in the context of digital game-based language learning programs was conducted by Govender and Arnedo-Moreno (2021). The most frequently analyzed programs were those for vocabulary learning, considering different age groups and target languages. According to the analysis, the most frequently used game design elements were feedback, theme, points, narrative, and level. The authors note, however, that generalizing results is usually difficult because no attention is paid to specific game design elements or their combinations. This problem is already known and researchers are looking for a systematic examination of individual game design elements to draw more precise conclusions about their effects (Mazarakis & Bräuer, 2018, 2022; Mekler et al., 2017).

There are numerous game design elements used in gamification research and practice, but the majority of gamification approaches are developed on a foundation of three game design elements: points, badges, and leaderboards (Huang et al., 2020; Seaborn & Fels, 2015; Werbach & Hunter, 2012). Based on the predominantly positive results for these three game design elements in many other studies, it was decided to investigate the effect of these elements in our study as well.

Points are most often the basis for other game design elements, as they can be used to capture and store various game metrics (Werbach & Hunter, 2012). They serve as a numerical unit to indicate progress (Seaborn & Fels, 2015) or as a numerical representation of rewards for performing certain activities (Vitkauskaitė & Gatautis, 2018). Unlike binary feedback (right or wrong), points can be used to indicate how close an answer is to the optimal solution and are therefore often used in learning systems (Kapp, 2012).

Badges and achievements are a form of feedback awarded to the user for completing certain tasks (Marczewski, 2018).

Badges are typically viewed as a visual representation of achievements (Antin & Churchill, 2011; Groening & Binnewies, 2019; Werbach & Hunter, 2012). Because badges, as previously noted, are among the most studied game design elements (Huang et al., 2020; Seaborn & Fels, 2015), they are considered in this study. At the same time, however, visual components should be avoided in the context of IVAs. Since the two terms are often used synonymously in the literature (Werbach & Hunter, 2012) and the effect of badges and achievements are comparable, the decision was made to use achievements in the study. According to Antin and Churchill (2011), achievements serve five functions: setting goals, giving instructions, allowing reputation to be valued, creating a group feeling, or serving as a status symbol. Groening and Binnewies (2019) examine the effect of achievements in detail. Their results show that achievements can increase performance and motivation. However, as with many other game design elements, the effect depends strongly on the design. Based on the results of their study, the authors recommend using achievements with a high degree of difficulty and in small numbers.

A leaderboard is another game design element and consists of an ordered display of scores against a specific success criterion, along with the names of users (Groening & Binnewies, 2019). Its purpose is to make simple comparisons (Zichermann & Cunningham, 2011), especially addressing the basic need for social relatedness (Bräuer & Mazarakis, 2019; Sailer et al., 2017).

Ortiz-Rojas et al. (2019) integrated a leaderboard into an engineering course and compared learning performance, intrinsic motivation, self-efficacy, and students' engagement with a control group that did not have a leaderboard over a four-week period. To assess learning performance, the results of a knowledge test were used. The other three variables were assessed using a questionnaire, which was completed at the start and end of the study. The integration of the leaderboard had a positive effect on learning performance, according to the findings. On the other hand, the other three factors did not become statistically significant. Finally, Landers and Landers (2014) investigated the impact of a leaderboard in the context of a course project. The authors assigned students to a leaderboard randomly and discovered that the leaderboard was associated with more time spent working on the project and, as a result, better project performance.

3. Method

After presenting the related literature, we now detail the methods in this chapter. The study examines the impact of audio-gamification on motivation and learning performance and thus investigates the integration of audio-implemented game design elements with IVAs. The hypotheses and experimental design are described in the following section, including a detailed description of the data cleaning for the statistical analysis.

3.1. Hypotheses

Based on previous findings in the scientific literature, we developed two hypotheses for the experiment, each split into comparisons between the control group and an experimental group. These are measured as the number of vocabulary processed, and learning success is measured as the number of words correctly reproduced.

The duration over which the subjects interacted with the experiment, more precisely with the Alexa skill, was analyzed to investigate the effect of the game design elements on user motivation. Because previous research has shown that gamification can increase general engagement (Landers & Landers, 2014; Majuri et al., 2018; Monterrat et al., 2015), and gamification encourages learners to engage more deeply with their learning material (Bouchrika et al., 2021; Dichev & Dicheva, 2017), the number of vocabulary processed was analyzed to determine whether the incorporation of game design elements influences learning motivation. As a result, we research the following two hypotheses:

H1a: The group with the leaderboard condition processed more vocabulary than the control group, with no game design elements.

H1b: The group with achievements condition processed more vocabulary than the control group, with no game design elements.

Second, the impact of the game design elements on subjects' learning performance will be investigated. Again, we refer to the literature results presented earlier (Denny et al., 2018). The ratio of vocabulary correctly processed is analyzed to measure learning success. Two additional hypotheses are examined:

H2a: The group with the leaderboard condition has a higher ratio of correct vocabulary than the control group, with no game design elements.

H2b: The group with the achievements condition has a higher ratio of correct vocabulary than the control group, with no game design elements.

3.2. Experimental design

A skill was developed for the Amazon Alexa platform that allows learning German as a foreign language to collect the data for this study. Because the speech input is determined by Alexa's default system language, this means that an IVA with an English setting as a system language cannot process German-language input. However, with the help of Amazon Alexa's Speech Synthesis Markup Language, it is also possible to reproduce texts in other languages *via* text-to-speech (*Speech Synthesis Markup Language (SSML) Reference | Alexa Skills Kit*, n.d.). Therefore, the Alexa skill is designed to train listening comprehension, according to the flashcard concept that Skidmore and Moore (2019) describe in their article. In addition to the advantage of not requiring users to make any additional system settings, this approach

provides the benefit of assuming that native speakers can easily interact with the IVA (Wu et al., 2020).

To ensure that a long-term effect can be demonstrated, more than 500 vocabularies were selected and integrated into the Alexa skill so that the users are always provided with new vocabularies, even when using the Alexa skill intensively. Of the words, 70% were nouns, 20% were verbs, and 10% were adjectives. In addition, many words regularly used in everyday language were chosen to make the Alexa skill enjoyable to beginners. Looking at the corresponding word frequency level, 38% of the words in the Alexa skill correspond to level 1, 25% to level 2, and only 4% to level 3. The remaining 33% of words could not be classified *via* the "Frequency Level Checker" tool to determine the word frequency level (Maeda, n.d.).

Some terms were frequently misunderstood when conducting pre-tests and had to be excluded. For example, "to buy" was misinterpreted as "goodbye," resulting in the Alexa skill being mistakenly ended in numerous incidents. However, the issue is widely recognized (Wu et al., 2020), and developers should always guarantee that every expected user input is unique to reduce user annoyance. This challenge is amplified when using Alexa skills to learn foreign languages because of the imprecise pronunciation of L2 learning (Wu et al., 2020). Nevertheless, the difficulty was mitigated in our study since the Alexa skill trains listening comprehension rather than pronunciation. Furthermore, the speech input was given in the trainees' native language, which in our instance was English. Still, Alexa may misinterpret even L1 speakers. Therefore, we counted how often each word was classified as correct or incorrect in the pre-tests. As a result, suspicious words were studied in-depth and replaced with less error-prone words.

Subsequently, to check the error susceptibility of the Alexa skill, an evaluation of the Automatic Speech Recognition (ASR) was performed using the test tool provided by Amazon. Using a test set created *via* the Alexa Developer Console, consisting of 762 test cases, a value of 75% was achieved. The ASR evaluation tool can be used to test audio samples with ASR models and compare the expected transcriptions with the actual transcriptions. The ASR analysis thus shows, for example, when "hettie" is understood instead of "headache." In addition, an intent confidence by utterance of 100% can be reported for the study period. No case was recorded with a low confidence, so according to the analysis, a proper answer was always found for the user's input.

To be able to investigate the influence of audio game design elements in a controlled setting, a 3×2 between-subjects design was implemented in which the control group was compared with each experimental group. Thus, the experimental setting includes the following groups:

- The control group without any game design element
- An experimental gamified group with a leaderboard
- An experimental gamified group with achievements

To acquire as many subjects as possible for the study, the Alexa skill was made available in all English language variants in the Amazon Store (US/CA/AU/IN/UK). Log files were used to record how often individual subjects used the Alexa skill, how long they interacted with it, how many vocabulary words were learned per session, and the success rate.

The experimental design of the Alexa skill differs among the groups. The two experimental conditions were each enhanced by the group-specific game design elements and functionalities to support gamification. The control group did not have any of these functionalities.

When subjects start the experiment for the first time, they are randomly assigned to one of the three conditions and cannot switch between the groups. They are informed that the Alexa skill is being used in a research project and that data will be stored for this purpose. The subject was also informed where to obtain more information about the Alexa skill and the research project. In addition, it was highlighted that the participation was entirely voluntary and that the deletion of the account and the associated data is possible at any time. The subject's anonymity was guaranteed by not asking for any additional information. This procedure is identical in all three groups. Any variations are described accordingly. In addition, section 3.3 provides a more detailed description of how and when the data were collected. After the introduction, vocabulary training begins. Figure 1 illustrates this procedure.

In the following, we first describe how the control group was designed and then go into more detail about the design of the individual game design elements for the two experimental conditions.

3.2.1. Design of the control group

In the control group, the vocabulary was asked on each trial until the subject answered all vocabulary or quit the Alexa skill whenever they wanted. The order in which the vocabulary is asked corresponds to the levels used in both other groups, which are explained in more detail in the next section. After each subject's response, either the correctness of the given answer is validated with positive oral feedback (e.g., "this is correct") or negative oral feedback, and the proper translation is provided. If a user has processed the entire vocabulary, learning will start from the beginning. The process is illustrated with a simplified flow chart in Figure 2. Unless the Alexa skill does not understand the user because (a) no input is provided or (b) an answer is given that the Alexa skill does not understand, Alexa apologizes and requests the answer to be repeated. This is repeated twice. If the Alexa skill does not receive an answer after the second request, Alexa switches off.

3.2.2. Design of the gamified groups

Both gamified groups (achievements and leaderboard) differ from the control group in some aspects. These are described in this section and additionally illustrated at the end of this section in Figure 3. In contrast to the control group, after informing the user that the data is being collected for research purposes, the subjects are asked for their nickname

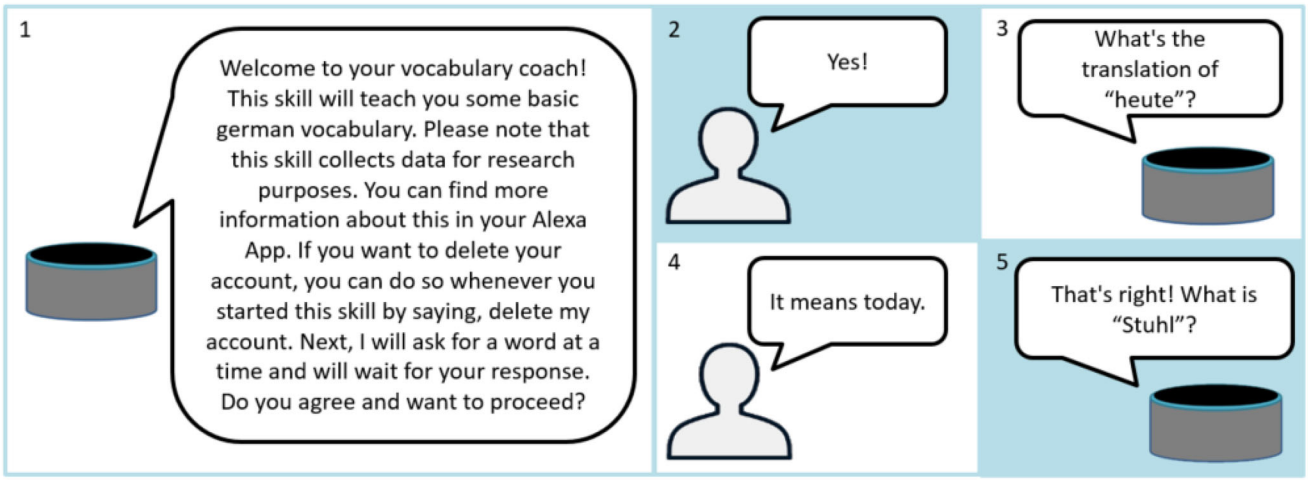

Figure 1. Example dialog in the control group when starting the Alexa skill for the first time.

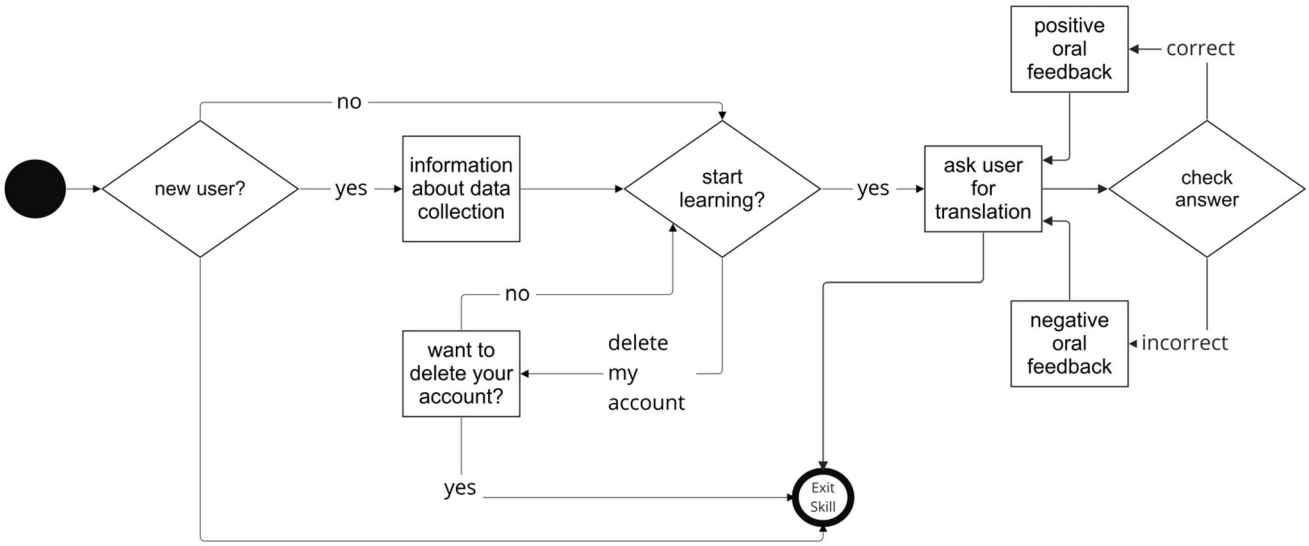

Figure 2. The flowchart illustrates the interaction with the Alexa skill in the control group.

at the first start. The nickname is used for a personal greeting when the experiment is accessed again. It is not necessary for the subject to state their real name; any nickname is acceptable. Now the user can choose between five topics: everyday life, health, leisure, nature, and education. The decision leads to the presentation of vocabulary from the chosen topic. In addition, the subjects are informed of their progress in percent. To avoid overwhelming the subjects with too much information and based on pre-test findings, the information about the progress only is given with a probability of 20%, so generally speaking, on average, every fifth time. Depending on the group, with leaderboard or achievements, the subjects also get information about the features of the respective game design element (see Figures 4 and 5).

The five topics are divided into 13 different levels, each topic consisting of 100 vocabulary words. The first six levels contain five different vocabulary words, and from the seventh level onwards, ten different vocabulary words are requested.

The subjects begin with the training session identical to the control group's vocabulary retrieval procedure, i.e., vocabulary is learned. After all vocabulary from a level has been asked and correctly answered at least once, the subject can practice the level again or take a challenge. During a challenge, all vocabulary from the corresponding level are randomly asked again.

When starting a challenge, the user receives life points. Each incorrect answer costs the subject one of their life points. If they run out of life points and the challenge has not been successfully completed yet, they lose and can restart the challenge from the beginning. If the subject chooses not to complete the challenge or fail it, there is the option to return to the main menu. Lower levels (1–6) begin with three life points, while higher levels (7–13) begin with four. After a challenge has been passed or failed, a corresponding audio feedback is heard, which is intended to emphasize the result of the challenge. An example of the dialog sequence in which a challenge is started is shown later in Figure 5.

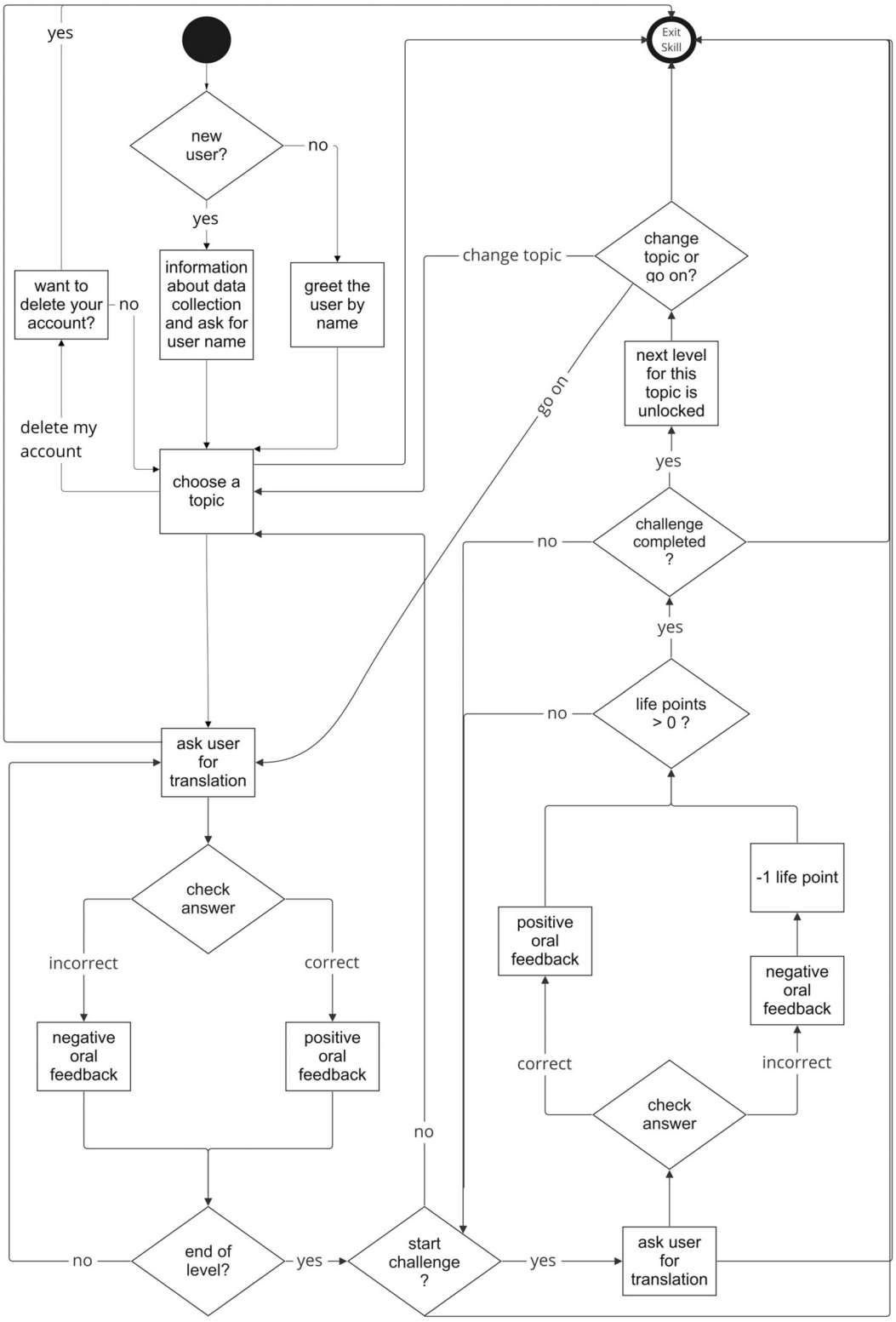

Figure 3. General flowchart illustrates the interaction with the Alexa skill in the experimental groups (achievements and leaderboard).

Once a challenge has been successfully completed, the next level for the corresponding topic is unlocked, and positive audio feedback is played. The subject may proceed directly to the next level's training unit or return to the main menu to change the topic. After completing all levels of a topic, the user gains access to a final challenge in which all vocabulary from all levels of the area is randomly asked.

3.2.3. Design of the leaderboard

The design for the first experimental group, the leaderboard condition, differs from the other groups. When the Alexa skill is activated, the subjects are told their position on the leaderboard with a 20% likelihood, according to the pre-test findings, to avoid overwhelming the subjects with too much information.

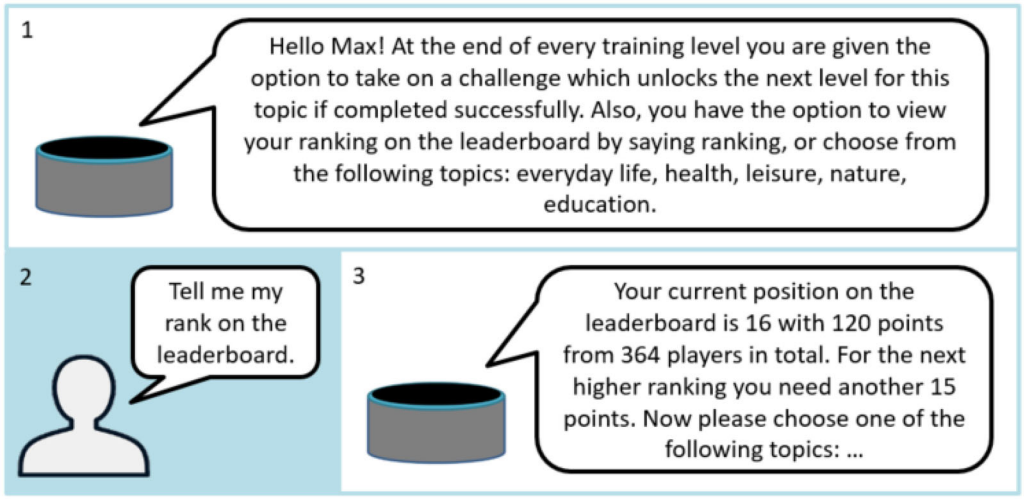

Figure 4. Example of a dialog in which the user calls up the leaderboard.

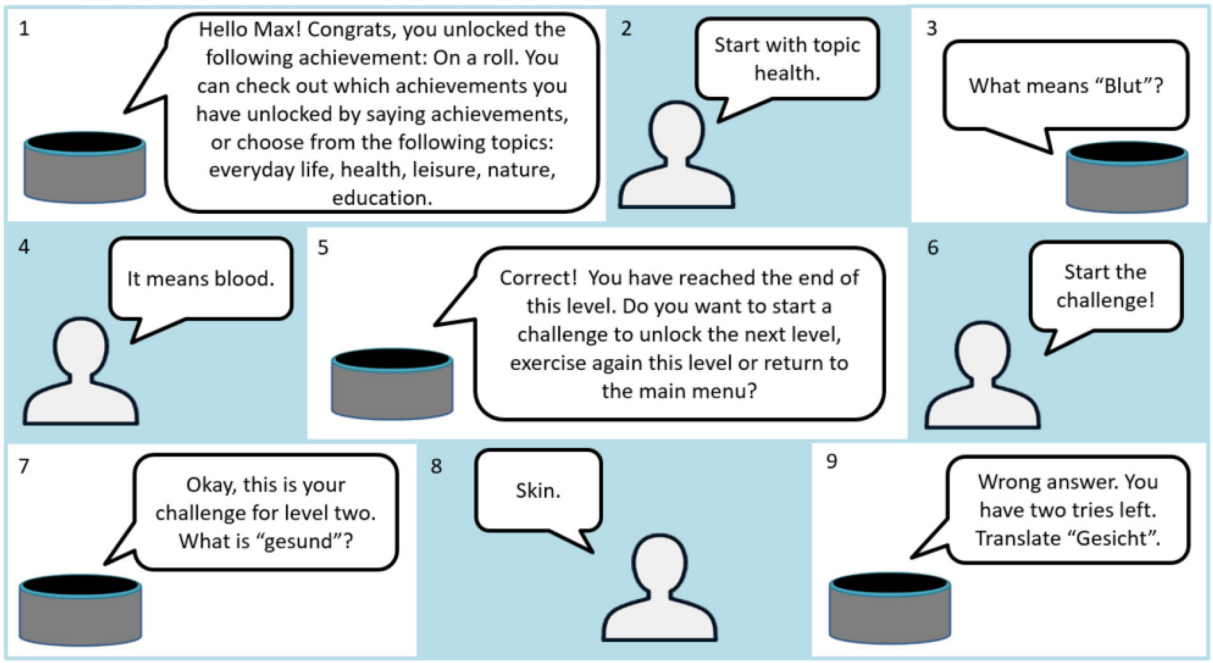

Figure 5. Example dialog for a user who receives an achievement and starts a challenge.

The subjects in this condition collect points during the challenges to move up in rank. The number of points awarded for a completed challenge depends on three factors: the level, whether a question was answered correctly, and how many times the challenge was started. For example, for levels 1–6, 125 points for each correct answer were awarded for the first attempt, 25 for the second, and five for the third and further attempts. The corresponding points for levels 7–13 were 250, 50, and 10. The staggered scoring is intended to encourage the user to complete a challenge in as few attempts and as good as possible, thus expediting their progress. Additionally, this method of awarding points is intended to prevent users from cheating. For example, a user earns no points by intentionally aborting and restarting the same challenge. An example of how the leaderboard is communicated to the user is shown in Figure 4.

3.2.4. Design of the achievements

Users in the second experimental group, the achievement condition, can unlock eleven achievements under certain conditions, as listed in Table 1. According to Werbach and Hunter (2012), badges or achievements can be used to assist in onboarding a new system. Badges should therefore be easy to reach right from the start so that users are encouraged to continue interacting with the application (Kapp, 2012). Therefore, the design of the achievements tries to give the user the possibility to unlock them already at the beginning, e.g., for completing a challenge for the first time or even giving a wrong answer. Another criterion for design is based on the results of Groening and Binnewies (2019). Therefore, the number of achievements was limited and was designed to be more challenging, such as working through a whole topic or reaching a final challenge.

Table 1. List of all achievements with their title and description.

Achievement	Description
Getting started	The user has completed the first challenge.
Every beginning is difficult	The user has failed a challenge.
Perfect one	The challenge of a ten-word level was completed without a mistake.
Halfway there	The user has completed fifty percent of all levels.
Topic master	The user has completed all the final challenges.
On a roll	Unlocked when the user has used the Alexa skill for three consecutive days.
Ambitious	The user has started the Alexa skill 20 times.
Practice makes perfect	Before starting a challenge, the user took the chance to practice the level again.
Time for the real challenge	The user has unlocked the first final challenge.
That is not what I meant, but okay	The user has used a registered synonym as an answer during training or during a challenge.
You are the vocabulary coach	The user has unlocked all the previous achievements.

An announcement informs subjects when an achievement is unlocked. These occur based on the state of the achievement at the end of a challenge, after answering a question, or when starting the Alexa skill. In addition, when an achievement is unlocked, audio feedback is played. The subjects can have their unlocked achievements read aloud to them from the main menu. The title of the achievement and its meaning are briefly explained. The user is then returned to the topic selection. An example of a dialog in which the user receives an achievement when starting the Alexa skill is shown in Figure 5.

3.3. Data collection

The underlying data for the results were collected in a long-term field experiment for one year, more precisely, between August 2020 and August 2021. For this purpose, the necessary Alexa skill was made freely available for download in the Amazon Store. The Alexa skill and the experiment were not advertised. Any participation was based on the random download of the Alexa skill from the store. During the experiment period, the Alexa skill was used by 488 unique users. No payment was offered or paid to the subjects.

Users were informed that their data was being gathered for research reasons on the Alexa skill's webpage in the Amazon Store and the Alexa app, as well as when they first started the Alexa skill. Each skill uploaded to the Amazon store automatically gets a webpage where users can see more information about that skill. This page can be found on the Amazon market, similar to pages for products sold on the platform. Our Alexa skill page explains what information is stored, such as the nickname and interaction logs. In addition to the website, each skill receives an entry in the Alexa app. The app is used to set up and operate Alexa devices. When a user activates a new Alexa skill, it appears in the app. Here, too, information about the research project and instructions on how to delete the collected data were provided for our Alexa skill. In addition, when starting the Alexa skill, subjects were told that data is stored for research purposes and that more detailed information is available in the Alexa app.

Before the subjects could continue interacting with the Alexa skill, they had to give consent by saying "yes" to the fact that they understood this information and wanted to continue using it. If a subject does not agree with this requirement at any time during the experiment, it is possible for them to delete the collected data *via* voice command. Because this step cannot be reversed, the user is asked to confirm this action twice before proceeding. It is possible to restart the experiment after deleting the data. However, the subject is treated as a new user and may be assigned to a different group due to the randomized assignment. The delete function was used four times throughout the experiment.

For the experiment, log files were collected that documented the usage behavior of the Alexa skill. For each user, an entry was created in a database when the Alexa skill was started for the first time. Interactions with the Alexa skill were then recorded, and the corresponding information, such as the correct or incorrect answer to a vocabulary word, was stored in the database. All data collected is anonymized, and user re-identification is not possible. To evaluate the hypotheses, the following general data were collected: ID of the user, group assignment, start and end time of a session, and frequency of use of the Alexa skill. The frequency of use increases with each access to the Alexa skill. The usage time is calculated from the start and end time entries in the database. In addition, the total number of vocabulary processed and the proportion of correctly and incorrectly answered vocabulary are stored.

In both gamified groups, additional information was stored to implement the game design elements. These include the nickname of the user, which level in which topic the user has reached, the total percentage of progress, the current score, the number of times the leaderboard has been accessed, and the current number of times a challenge of a specific level has been taken. In addition, for the group with achievements, the unlocked achievements are saved with the timestamp when they got unlocked and the number of times the achievement overview has been accessed.

3.4. Participants

Not all data collected during the study period can be used for the subsequent analysis of the hypotheses. This chapter, therefore, describes the procedure used to exclude data sets that were not usable for the analysis. Subsequently, a

Table 2. Mean, standard derivation, and median of the sum of processed vocabulary, correct answers, and time in seconds per user.

	Sum vocabulary			Correct vocabulary			Time in seconds		
	M	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Control	24.68	36.85	9.50	11.30	20.82	2.00	287.83	367.97	144.50
Achievements	25.79	30.49	14.50	13.23	17.45	7.00	418.58	393.96	275.00
Leaderboard	26.87	34.44	16.50	16.12	25.15	9.00	392.14	379.29	267.50
Total	25.86	34.13	13.00	13.75	21.88	5.00	364.80	381.98	223.50

description of the subjects is summarized based on the cleaned data.

3.4.1. Data cleaning

Data cleaning is performed based on two data sets: The sessions conducted and the subjects. A session has a starting point when individuals start the Alexa Skill and an endpoint when they stop using the Alexa skill, intentionally or unintentionally. A new session is created when a user invokes the Alexa skill a second time. To merge the data of the individual sessions, a unique user ID generated by Amazon is used. This unique key remains the same even if the user calls up the Alexa skill again so that the assignment to the conditions remains the same and the progress in learning the vocabulary can be documented. Over the entire period of the study, data was collected on 488 subjects (156 in the control group, 152 in the achievements group, and 180 in the leaderboard condition) with associated data on 941 sessions (individual access to the Alexa skill).

First, from 488 unique user IDs in the dataset, 229 entries were deleted because not a single vocabulary was processed over all sessions. In addition, another 17 subjects who had listened to only one vocabulary word but answered it incorrectly were removed. In the next step, one subject was removed from the analysis because the behavior was characteristic of a bot and not a human (access always took place at the same time, and none of the asked vocabulary words were answered correctly over 84 sessions). Finally, another subject was excluded from the dataset because out of 83 vocabulary words, zero were answered correctly. After these steps, entries from 240 subjects who interacted with the Alexa skill in 375 sessions remain.

Finally, outliers were excluded from the dataset because they studied a disproportionately large number of vocabulary items and appeared to be highly intrinsically motivated. The mean value of the processed vocabulary for all users (40.70) was calculated for this purpose, and the double standard deviation (2* 85.40) was added. This led to removing ten more users from the dataset who edited more than 211.50 vocabulary words. The final sample for analysis consists of 230 subjects and 289 combined sessions.

3.4.2. Description of the final sample

The 230 subjects are distributed among the three experimental conditions, with 76 in the control group, 62 in the achievements group, and 92 in the leaderboard group. Besides, the Alexa Developer Console revealed which language settings were used on the devices by users: 24% from the United Kingdom, 5% from Canada, 46% from the United States, 2% from Australia, and 23% from India. While this enables conclusions about users' locations to be drawn, it is not the same as geolocations. For example, while a user in Canada is more likely to set the language to Canadian English, they could just as easily set it to British English or even German for their end device.

4. Results

This section provides the descriptive and inferential statistical results. This includes the effect of gamification on usage behavior and on learning success. Finally, further statistical analysis of whether the assumptions regarding the time duration are valid is conducted.

4.1. Descriptive results

In the control group, the experiment was started on average 1.29 times per user (*SD* 0.78), in the group with achievements 1.35 times per user (*SD* 0.89), and in the group with leaderboards 1.16 times per user (*SD* 0.50). So for repeated Alexa skill usage, no noticeable statistical differences or biases exist between the groups.

To answer the hypotheses, the means, standard deviations, and medians for the total vocabulary processed, the percentage of vocabulary words answered correctly, and the time in seconds that users interacted with the Alexa skill are summarized in Table 2.

4.2. Effect of gamification on usage behavior

To evaluate the hypotheses, we first examine how the game design elements affect the number of vocabulary processed. The data are first tested for normal distribution before comparing the mean values between the control and experimental groups. The Shapiro-Wilk test produces a statistically significant result, $p < .001$. As a result, the Mann-Whitney U test is used as a non-parametric statistical method for comparing mean values between groups (Field, 2017).

The comparison between the control group and the experimental group with achievements is not statistically significant: $U = 2002$, $z = -1.52$, $p = .065$ (one-tailed), $r = .13$. As a result, hypothesis H1a cannot be supported, and it cannot be assumed that achievements have a positive impact on learners' motivation to engage more deeply with their learning material.

The comparison between the control group and the group with the leaderboard yields a statistically significant result: $U = 2873$, $z = -1.99$, $p = .024$ (one-tailed), with a small effect size $r = .15$. The effect size for the hypothesis

Table 3. Mean, standard derivation, and median ratio of correct vocabulary to the sum of answered vocabulary.

	M	*SD*	*Mdn*
Control	0.300	0.229	0.333
Achievements	0.398	0.265	0.417
Leaderboard	0.457	0.253	0.500
Total	0.390	0.257	0.417

H1b is interpreted according to Field (2017). Thus, hypothesis H1b can be supported, and it can be assumed that adding a leaderboard increases the motivation to practice more vocabulary with the Alexa skill.

4.3. Effect of gamification on learning success

The ratio of correct vocabulary to the sum of answered vocabulary was first calculated to investigate the impact of game design elements on learning success. This is necessary to answer the hypotheses H2a and H2b. Table 3 summarizes the findings. For hypotheses H2a and H2b, the interpretation of the effect size is made according to the benchmarks provided by Plonsky and Oswald (2014) for L2 research.

The sample was tested once more for normal distribution for further analysis. The Shapiro-Wilk test also becomes statistically significant here, with $p < .001$. The comparison between the control group and the experimental group with achievements becomes statistically significant, $U = 1829.5$, $z = -2.27$, $p = .012$ (one-tailed), with a small effect size of $r = .19$. In L2 research, a correlation coefficient interpreted as effect size has a small effect size around 0.25 and a medium effect size around 0.40, leaving a large effect size around 0.60 (Plonsky & Oswald, 2014). As a result, Hypothesis H2a can be supported, and it can be assumed that achievements have a positive influence on learning success.

The comparison between the control group and the experimental group with a leaderboard gives a statistically significant result as well, $U = 2223.5$, $z = -4.07$, $p = .000$ (one-tailed), with a small effect size of $r = .31$. Thus, the hypothesis H2b can also be confirmed, and it is assumed that leaderboards also positively influence learning success.

4.4. Further analyses

The time users spend with the Alexa skill per session is another aspect that could yield further interesting insights about the influence of audio-gamification. Since the dialogues in the gamified groups are much longer than in the control group (as can be seen by comparing Figures 2 and 3), no firm conclusions about learning behavior can be established. However, time per session may still provide insights into user behavior. In line with the previous hypotheses, it is assumed that in the gamified conditions, there are increases compared to the control group. Following that, we examine how the game design elements affected the interaction time of use in seconds. Again, we first check the normal distribution of the data and obtain a statistically significant result using the Shapiro-Wilk test, $p < .001$. As before, we need to use the Mann–Whitney *U* test, a non-parametric statistical method for comparing mean values between groups (Field, 2017).

The comparison between the control group and the experimental group with achievements becomes statistically significant, $U = 1495$, $z = -3.68$, $p = .000$ (one-tailed), with a medium effect size $r = .31$. Furthermore, the comparison between the control group and the experimental group with the leaderboard condition provides a statistically significant result, $U = 2257$, $z = -3.95$, $p = .000$ (one-tailed), also with a medium effect size $r = .30$.

Prior knowledge of the subjects could bring a bias into the data in such a field experiment, which could significantly complicate the interpretation of the results. For reasons of data economy and following the privacy by design principle, we did not ask for this information in our study. However, the analysis of the existing data is sufficient to exclude such a bias. The main idea for identifying such a bias lies in the skewness and distribution of data. Despite the premise that, in general, a sample size of 30 and more subjects, especially if it is the case in each group, is widely accepted as large enough (Field, 2017), we can still analyze the data further to find irregularities about learning behavior. For this reason, we further analyzed the ratio of correct vocabulary. First, all students with zero correct answers were removed, as prior knowledge does not exist for them and cannot lead to a bias. The number of subjects removed is almost equally distributed across all three groups (20 for the control group and 12 each for the achievements and leaderboard groups). Next, it was analyzed if the data distribution and skewness of the data for the remaining subjects in each group were normal. This is indeed the case: The skewness for the control, achievements, and leaderboard groups are 0.024 (*SD* 0.319), −0.102 (*SD* 0.337), and −0.277 (*SD* 0.269), all far away from a critical value of 1 or −1 (Field, 2017). In addition, the Shapiro-Wilk test is not statistically significant for any group, with $p = .131$ for the control group, $p = .424$ for the achievements group, and $p = .235$ for the leaderboard group. These results indicate that there is no apparent bias in the groups and that the participants have no obvious bias in their learning behavior. Of course, this is not strong statistical evidence that the participants are similar in terms of learning attitudes. Nevertheless, it is still a powerful indicator that the subjects in each group do not differ in learning behavior, which could be the case if participants in one group would benefit from prior knowledge.

5. Discussion, limitations, and future research

The motivation for this long-term field experiment was to advance the field of conversational user interfaces with the combination of incentives to use and so to enhance language learning with an IVA. Therefore, we systematically defined, analyzed, and compared the use of audio-gamification for one year with a field experiment on practicing listening comprehension of German vocabulary with an Amazon Alexa skill.

What distinguishes this study from previous research in both (a) IVAs for language learning (Dizon, 2021) and (b)

gamification for language learning (Dehghanzadeh et al., 2021; Huang et al., 2020; Seaborn & Fels, 2015) is the combination of IVAs and gamification as well as the collection of data *via* a long-term field experiment. By conducting the study in a natural non-school setting, i.e., informal and non-guided, realistic findings on the use of IVAs as a tool for L2 learning could be collected (Dizon, 2021) and thus achieve a high external validity. At the same time, in contrast to many other studies (Dehghanzadeh et al., 2021), this article also reports the results of a control group, which distinguishes it from many other studies. In line with previous literature on gamification, we find positive effects but also mixed results.

Regarding the first hypothesis, a statistically significant influence on the number of processed vocabulary words was found using the leaderboard but not for the achievements condition. As already briefly described, in contrast to the visual representation of gamification, every game design element in audio-gamification must be represented by sound or speech. Consequently, information about unlocking achievements cannot be displayed as, e.g., a small element in the upper right corner of a screen. Instead, the actual activity must be interrupted for the game design element to be expressed. As a result, the game design elements are at odds with increasing motivation and distracting from the actual activity. Since the subjects in the group with achievements interact longer with the Alexa skill but do not automatically process more vocabulary, such a conflict is possible in general and also in the experiment. As an implication, we recommend investigating how much additional time is needed for the presentation of the game design elements in the design process, and if it interferes with the motivation of the individuals.

Another aspect that may explain why the achievements do not contribute to the processing of more vocabulary can be attributed to their design in the study. Users could access their achievements, but there was no way to make them available to other users. What makes an achievement desirable is if it is available to a larger public and whether or not it is seen by others (Zichermann & Cunningham, 2011). It is more challenging to motivate users to collect achievements without comparison. Future studies should investigate if and how it is possible to access the game design element achievements for others.

Another design aspect that may have worked against the desired goal and thus weakened the effect of the achievements is the use of achievements for negative events. The goal in designing the achievements was to keep the subjects interested by offering a wide variety of unlocking options. Unlike the rest, two achievements, "every beginning is difficult" and "that is not what I meant, but okay" were unlocked when an answer was not quite appropriate or completely wrong. By awarding such negative achievements, the user is explicitly made aware of his mistake, which can have a demotivating effect (Kapp, 2012).

Perhaps achievements (and badges) are not suitable for a purely audio-based implementation. Since we based the selection of game design elements on those most intensively considered in gamification research (Christy & Fox, 2014; Sailer & Sailer, 2021; Tan & Hew, 2016), which are usually implemented visually, it is not unlikely that these elements will not work in a purely audio-based environment (Friberg & Gärdenfors, 2004). To identify more appropriate game design elements for audio-gamification, an analysis of audio games would be helpful. These elements could then be applied to a similar context as in our study.

As for the second hypothesis, a statistically significant effect of both game design elements on learning outcomes was obtained. This result also supports the findings of previous research on gamification in education (Denny et al., 2018; Sailer & Homner, 2020). Thus, audio-gamification offers the potential to enhance further positive effects on learning outcomes through IVAs, such as those obtained by Dizon (2020).

Contrary to the results of Dizon (2020), and Hsu et al. (2021), a positive effect on the improvement of listening comprehension was found in the study. The different conditions of the studies provide a possible explanation for this. In both experiments, Dizon (2020) and Hsu et al. (2021) subjects trained their listening comprehension by interacting with an IVA whose system language was set to the foreign language being learned. In our study, the basic integration with the IVA occurred in the user's preselected language, which can be assumed to be the native language. Only the vocabulary words to be learned, which were asked one by one, had to be understood correctly. Dizon (2020) assumes that the lack of effect is possibly due to the excessively high speaking rate and the too difficult vocabulary of the IVA. Both factors could be reduced in the study setting.

The evaluation of the interaction time, which was carried out in addition to the hypotheses, also differed statistically between the groups. Thus, the results could support the conclusions of previous studies on the effect of gamification (Landers & Landers, 2014; Majuri et al., 2018; Monterrat et al., 2015). Furthermore, we confirm that this potential can also be transferred to the context of IVAs and that the positive effect can also be generated by purely audio-based game design elements. As already proven in a laboratory experiment before (Bräuer & Mazarakis, 2022), audio-gamification can increase user engagement with an application. We were able to reproduce these results in a real-life situation. Unlike Bräuer and Mazarakis (2022), where several game design elements were used at once, this study goes into detail about the effect of the two individual game design elements examined, i.e., achievements and leaderboards. Our results take on the findings of the earlier study further by showing that audio-gamification works, but not every successful gamification concept can be readily transferred to the IVA context. However, the significance of these results should be interpreted with caution since the dialogs in the gamified groups are significantly longer than in the control group. Nevertheless, it is interesting that the subjects were not tempted by Alexa's more complex instructions to work on fewer vocabulary items, e.g., to always stop after 5 min regardless of which group they were in.

Despite the positive results of the study, when integrating audio-gamification into applications for IVAs, it should be

noted that its integration increases the complexity of the application. Luger and Sellen (2016) point out that to interact with an IVA, users need time to grasp the scope of interaction options. Unlike visual interfaces, which allow users to see numerous options simultaneously, an IVA, similar to an audio game, requires each option to be introduced by speech (Friberg & Gärdenfors, 2004). For example, the category selection was always read aloud completely, which would not be necessary for a similar program on a device with a screen. Because the incorporation of game design elements might enhance the complexity of the conversational user interface, the time required for users to become acquainted with an application is likely to grow. As a result, it should constantly be verified whether gamifying an application makes it more difficult to use.

In this study, users could receive an overview of their achievements or their current position on the leaderboard in the main menu. However, these features were only used three times by the subjects. On the one hand, this very low interaction with the overview-features could be interpreted as very little engagement with the gamification system. Nevertheless, this is not the case because even without actively calling up these features, the users were constantly made aware of their progress through the integrated game design elements. For successful gamified systems, the user-system interaction plays a crucial role (Liu et al., 2017). User-system interactions include system-user communications (Liu et al., 2017, p. 1015) and aspects which provide feedback and leaderboards to monitor progress (Liu et al., 2017, p. 1023). This is the case for our experiment. As the positive effect of the leaderboard condition shows, calling up the overview does not seem necessary to increase motivation. From a design perspective, it should be reconsidered whether the integration of such an overview is necessary or makes the system unnecessarily complex.

The dropout rate of 47% of the subjects who accessed the experiment without processing even one vocabulary is worth mentioning. These individuals probably accessed the Alexa skill by mistake or lost interest in the Alexa skill directly after the explanation at the beginning, e.g., when the statement came that data is recorded for research purposes. Privacy concerns could potentially be a reason for direct termination in this scenario. Because many users are suspicious about an IVA processing personal data (Lau et al., 2018; Liao et al., 2019), the explanation about data gathering for research objectives when starting the Alexa skill may dissuade these users. However, from the standpoint of research ethics, such an explanation should not be missing. In addition, also other studies report high dropout rates, e.g., 38% in an online experiment about visual preferences (Jun et al., 2017). So, although we had anticipated a lower rate, this relatively high rate is not a cause for concern. Still, 230 subjects could be analyzed inferentially to test the hypotheses.

In addition to the high rate of users who did not use the Alexa skill at all, it should also be noted that many only used the Alexa skill once. Overall, 37 subjects, or 16% of the users in the sample for analysis, used the Alexa skill more than once. Nevertheless, although this number seems low initially, it is relatively high in comparison. This is because the recurrence rate of applications for IVAs is generally very low and currently lies at 6% on average (Planner, 2018). In comparison, the Alexa skill developed for this study was thus activated more frequently, but the audio-gamification still does not seem to have really motivated regular use.

However, to learn a foreign language with the IVA in the long run, it would be necessary to use the Alexa skill over a longer period of time. One reason for the low multiple uses of the Alexa skill could be the relatively low mean percentage of correctly answered questions of 39%. The application might have been designed too difficult. The Alexa skill lacked a sequence in which the vocabulary was translated once before it had to be answered for the first time. To promote learning success, follow-up studies should first include a practice sequence in which all vocabulary words at the current level are presented once with the appropriate translation.

Also, to the design challenge of implementing audio-gamification, providing a foreign language learning Alexa skill brings some pitfalls that should not go unmentioned. When pre-testing the Alexa skill, we sort out as many words as possible that the IVA could potentially misunderstand. Nevertheless, individual inputs may not be correctly understood. Such a misunderstanding could certainly harm the user's motivation. In the field experiment, we did not store the user's input, but only recorded *via* a counter how often a vocabulary item was classified as correct or incorrect. To be able to exclude the influence of such errors on the user's motivation in future studies, a manipulation check in the form of precise storage of the input should be performed.

The goal of current gamification research should be to demonstrate the effect of individual game design elements so that specific conclusions can be drawn (Govender & Arnedo-Moreno, 2021; Landers, 2014; Mazarakis, 2021; Mekler et al., 2017). The game design elements achievements and leaderboard were considered in two different experimental conditions in the study. However, both featured several game design elements (points, levels, and challenges). Thus, each element's effect could have been strengthened or weakened by combining elements. The findings allow general assumptions about the effects of audio-gamification, but a detailed analysis of the effects of achievements and leaderboards is only possible to a limited extent.

As mentioned earlier, user fear of personal data processing by an IVA is a challenge (Lau et al., 2018; Liao et al., 2019). Therefore, to avoid discouraging users, the study refrained from asking for demographic data about the subjects, such as age and gender. However, previous research has shown that factors, such as age and gender can influence the effect of gamification (Codish & Ravid, 2017; Jent & Janneck, 2018). Similarly, users' age or gender can influence their IVA usage behavior (Zellou et al., 2021). Therefore, demographic data are vital for contextualizing the results, which might occasionally yield different interpretations than initially believed. In addition, the language level of the subjects, which was not asked in this study, can be necessary for data interpretation. For example, it is possible that some

demographic groups, such as women or persons with better language skills, were simply overrepresented in particular groups, which could potentially explain why these groups differed significantly. Additional information about subjects would be helpful in subsequent studies to uncover such effects, although users may be deterred by these questions for privacy reasons. It should be noted, however, that there is no evidence in the field experiment that the biases described here may be present.

As a study by Leung et al. (2022) showed, the effect of gamification is not only influenced by the selection of different game design elements. Adaptation to individual characteristics is also relevant, where different game design elements can cause different effects on online learning. One aspect that can influence the effect of gamified learning systems, according to Leung et al. (2022), is the individual goal orientation of the learners. Therefore, they recommend applying personalized approaches to gamification in the field of education in the future. To optimally design the effect of audio-gamification in the field of vocabulary learning, it would therefore also be worth considering a personalized approach to gamification for future studies, which is already acknowledged as an important topic for gamification in general (Mazarakis, 2021).

Another limitation of the study is that it did not record what type of device the users were using. Dizon et al. (2022) found that in a study with an Echo Show, unlike previous studies with a speaker without a display, new problems arose during the first interaction with the device. For example, users attempted to type their responses *via* the display instead of giving a verbal response. Even though in a field experiment, it is highly likely that users already have experience with IVAs, especially the time users spend with the Alexa skill could be influenced by such issues. To counteract this, care was taken during implementation not to show any information on the display. Only the name of the Alexa skill was displayed.

Although SDT is often used only to distinguish between intrinsic and extrinsic motivation (Koivisto & Hamari, 2019), SDT offers a much broader range of explanatory possibilities. For example, competitive game design elements may be perceived as introjected regulation of extrinsic motivation and have a favorable short-term influence in various fields. This complexity is already acknowledged (Tyack & Mekler, 2020), and future studies should use the full potential of the self-determination theory.

As Ruan et al. (2019) already showed, learning vocabulary *via* voice-based systems is time-consuming but, at the same time, seems to be more motivating than classical learning with article. The gamification integration, as done in our study, definitely increases the interaction time. Speech was mainly used, using audio feedback only when a challenge was completed or an achievement was acquired. However, when looking at audio games, music and sound are common to represent the mechanics of the game (Cicció & Quesada, 2018). So, for example, more diverse sounds or melodies could be used after unlocking an achievement. Similarly, the oral feedback that users receive when answering vocabulary could be replaced by audio feedback, as is often found in apps and games (Thiebes et al., 2014). In the field of audio game development, Friberg and Gärdenfors (2004) recommend providing individual sounds with several layers of information. For example, a sound can be used to indicate whether a machine is running or switched off and simultaneously convey how far away it is from the player. One recommendation that can be derived from our experience is to look for ways to integrate game design elements so that they can also be presented in parallel to the actual application. In this way, it might be possible to take advantage of gamification in the context of IVAs without losing the user's attention by making the interaction time too long.

6. Conclusion

In this long-term field experiment, quantitative data on the usage behavior of a gamified Amazon Alexa skill was analyzed. The research question is answered how the combination of IVAs and audio-gamification in the field of foreign language learning influences motivation and learning success. Using a field experiment instead of a lab experiment provides robust quantitative data in the area of conversational interfaces, in contrast to the primarily conducted qualitative studies. Data from 230 subjects were analyzed for one year. This revealed insights into design approaches and their effects on subject engagement and learning success. Thus, building on this study, the first conclusions can be drawn about which game design elements are particularly suitable for learning a new language in the context of voice assistants, e.g., leaderboards. This study provides results that contribute to opening new design perspectives for IVA research, as well as highlighting the aspect of audio elements from the perspective of gamification research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Paula Bräuer http://orcid.org/0000-0001-5903-8829
Athanasios Mazarakis http://orcid.org/0000-0001-9943-0382

References

Adams, S. (2019, August 31). *Game of tongues: How duolingo built a $700 million business with its addictive language-learning app*. Forbes. https://www.forbes.com/sites/susanadams/2019/07/16/game-of-tongues-how-duolingo-built-a-700-million-business-with-its-addictive-language-learning-app/

Antin, J., & Churchill, E. F. (2011). Badges in social media: A social psychological perspective. In *CHI 2011 Gamification Workshop Proceedings* (pp. 1–4).

Babic, S., Orehovacki, T., & Etinger, D. (2018). Perceived user experience and performance of intelligent personal assistants employed in higher education settings. In *2018 41st International Convention on Information and Communication Technology, Electronics and*

Microelectronics (MIPRO) (pp. 830–834). https://doi.org/10.23919/MIPRO.2018.8400153

Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, *30*, 100322. https://doi.org/10.1016/j.edurev.2020.100322

Benner, D., Schöbel, S., Süess, C., Baechle, V., Janson, A. (2022). Level-up your learning – Introducing a framework for gamified educational conversational agents. In *Wirtschaftsinformatik 2022 Proceedings*. https://aisel.aisnet.org/wi2022/hci/hci/5

Bouchrika, I., Harrati, N., Wanick, V., & Wills, G. (2021). Exploring the impact of gamification on student engagement and involvement with E-learning systems. *Interactive Learning Environments*, *29*(8), 1244–1257. https://doi.org/10.1080/10494820.2019.1623267

Bräuer, P., & Mazarakis, A. (2019). Badges or a leaderboard? How to gamify an augmented reality warehouse setting. In J. Koivisto & Hamari (Eds.), *Proceedings of the 3rd International GamiFIN Conference—GamiFIN 2019* (Vol. 2359, pp. 229–240). CEUR-WS.org.

Bräuer, P., & Mazarakis, A. (2022). "Alexa, can we design gamification without a screen?"—Implementing cooperative and competitive audio-gamification for intelligent virtual assistants. *Computers in Human Behavior*, *135*, 107362. https://doi.org/10.1016/j.chb.2022.107362

Cermak-Sassenrath, D. (2019). Current challenges in gamification identified in empirical studies. In R. Ørngreen, M. Buhl, & B. Meyer (Eds.), *Proceedings of the 18th European Conference on e-Learning (ECEL)* (pp. 119–127). Academic Conferences and Publishing International Limited.

Change the language of Google Assistant—Android—Google Nest Help (2021). https://support.google.com/googlenest/answer/7550584?hl=en&co=GENIE.Platform%3DAndroid#zippy=%2Cgoogle-nest-mini-nd-gen

Chen, H. H.-J., Yang, C. T.-Y., & Lai, K. K.-W. (2020). Investigating college EFL learners' perceptions toward the use of Google Assistant for foreign language learning. *Interactive Learning Environments*, 1–16. https://doi.org/10.1080/10494820.2020.1833043

Christy, K. R., & Fox, J. (2014). Leaderboards in a virtual classroom: A test of stereotype threat and social comparison explanations for women's math performance. *Computers & Education*, *78*, 66–77. https://doi.org/10.1016/j.compedu.2014.05.005

Cicció, J. A., & Quesada, L. (2018). Framework for creating audio games for intelligent personal assistants. In T. Ahram & C. Falcão (Eds.), *Advances in human factors in wearable technologies and game design* (pp. 204–214). Springer International Publishing. https://doi.org/10.1007/978-3-319-60639-2_21

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., Edwards, J., & R Cowan, B. (2019). The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, *31*(4), 349–371. https://doi.org/10.1093/iwc/iwz016

Codish, D., & Ravid, G. (2017). Gender moderation in gamification: Does one size fit all? In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2017.244

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., & Bandeira, N. (2017). "What can I help you with?": Infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human–Computer Interaction with Mobile Devices and Services* (pp. 1–12). https://doi.org/10.1145/3098279.3098539

de Barcelos Silva, A., Gomes, M. M., da Costa, C. A., da Rosa Righi, R., Barbosa, J. L. V., Pessin, G., De Doncker, G., & Federizzi, G. (2020). Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, *147*, 113193. https://doi.org/10.1016/j.eswa.2020.113193

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

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

Denny, P., McDonald, F., Empson, R., Kelly, P., & Petersen, A. (2018). Empirical support for a causal relationship between gamification and learning outcomes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3173574.3173885

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

Develop Skills in Multiple Languages | Alexa Skills Kit (2021). Amazon (Alexa). https://developer.amazon.com/de-DE/docs/alexa/custom-skills/develop-skills-in-multiple-languages.html

Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: A critical review. *International Journal of Educational Technology in Higher Education*, *14*(1), 9. https://doi.org/10.1186/s41239-017-0042-5

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Educational Technology & Society*, *18*(3), 75–88.

Dizon, G. (2017). Using intelligent personal assistants for second language learning: A case study of Alexa. *TESOL Journal*, *8*(4), 811–830. https://doi.org/10.1002/tesj.353

Dizon, G. (2020). Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology*, *24*(1), 16–26. http://hdl.handle.net/10125/44705

Dizon, G. (2021). Affordances and constraints of intelligent personal assistants for second-language learning. *RELC Journal*, 003368822110205. https://doi.org/10.1177/00336882211020548

Dizon, G., Tang, D., & Yamamoto, Y. (2022). A case study of using Alexa for out-of-class, self-directed Japanese language learning. *Computers and Education: Artificial Intelligence*, *3*, 100088. https://doi.org/10.1016/j.caeai.2022.100088

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Friberg, J., & Gärdenfors, D. (2004). Audio games: New perspectives on game audio. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology – ACE'04* (pp. 148–154). https://doi.org/10.1145/1067343.1067361

Garcia, F. E., de Almeida Neris, V. P., Garcia, F. E., & de Almeida Neris, V. P. (2013). Design guidelines for audio games. In M. Kurosu & M. Kurosu (Eds.), *Proceedings of the 15th International Conference on Human–Computer Interaction* (Vol. 8005, pp. 229–238). Springer. https://doi.org/10.1007/978-3-642-39262-7_26

Govender, T., & Arnedo-Moreno, J. (2021). An analysis of game design elements used in digital game-based language learning. *Sustainability*, *13*(12), 6679. https://doi.org/10.3390/su13126679

Groening, C., & Binnewies, C. (2019). "Achievement unlocked!"—The impact of digital achievements as a gamification element on motivation and performance. *Computers in Human Behavior*, *97*, 151–166. https://doi.org/10.1016/j.chb.2019.02.026

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? – A literature review of empirical studies on gamification. In *47th Hawaii International Conference on System Sciences (HICSS)* (pp. 3025–3034). https://doi.org/10.1109/HICSS.2014.377

Hsu, H.-L., Chen, H. H.-J., & Todd, A. G. (2021). Investigating the impact of the Amazon Alexa on the development of L2 listening and speaking skills. *Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2021.2016864

Huang, R., Ritzhaupt, A. D., Sommer, M., Zhu, J., Stephen, A., Valle, N., Hampton, J., & Li, J. (2020). The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational Technology Research and Development*, *68*(4), 1875–1901. https://doi.org/10.1007/s11423-020-09807-z

IOS and iPadOS-Feature Availability (2021). Apple. https://www.apple.com/ios/feature-availability/

Istrate, A. M. (2019). *The impact of the virtual assistant (VA) on language classes* (1.0, 163840. B) [Data set]. ADLRO. https://doi.org/10.12753/2066-026X-19-040

Jent, S., & Janneck, M. (2018). Using gamification to enhance user motivation: The influence of gender and age. In L. E. Freund & W. Cellary (Eds.), *Advances in the human side of service engineering* (Vol. 601, pp. 3–10). Springer International Publishing. https://doi.org/10.1007/978-3-319-60486-2_1

Jun, E., Hsieh, G., & Reinecke, K. (2017). Types of motivation affect study selection, attention, and dropouts in online experiments. *Proceedings of the ACM on Human–Computer Interaction*, *1*(CSCW), 1–15. https://doi.org/10.1145/3134691

Kapp, K. M. (2012). *The gamification of learning and instruction: Game-based methods and strategies for training and education*. John Wiley & Sons.

Kinsella, B. (2019, March 7). *U.S. smart display user base grew 558% in 2018 and more than doubled in second half of the year, Amazon holds two-thirds market share*. Voicebot.Ai. https://voicebot.ai/2019/03/07/u-s-smart-display-user-base-grew-558-in-2018-and-more-than-doubled-in-second-half-of-the-year-amazon-holds-two-thirds-market-share/

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, *45*, 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

La Guardia, J. G., Ryan, R. M., Couchman, C. E., & Deci, E. L. (2000). Within-person variation in security of attachment: A self-determination theory perspective on attachment, need fulfillment, and well-being. *Journal of Personality and Social Psychology*, *79*(3), 367–384. https://doi.org/10.1037/0022-3514.79.3.367

Landers, R. N. (2014). Developing a theory of gamified learning: linking serious games and gamification of learning. *Simulation & Gaming*, *45*(6), 752–768. https://doi.org/10.1177/1046878114563660

Landers, R. N., & Landers, A. K. (2014). An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance. *Simulation & Gaming*, *45*(6), 769–785. https://doi.org/10.1177/1046878114563662

Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human–Computer Interaction*, *2*(CSCW), 1–31. https://doi.org/10.1145/3274371

Legaki, N.-Z., Xi, N., Hamari, J., Karpouzis, K., & Assimakopoulos, V. (2020). The effect of challenge-based gamification on learning: An experiment in the context of statistics education. *International Journal of Human–Computer Studies*, *144*, 102496. https://doi.org/10.1016/j.ijhcs.2020.102496

Leung, A. C. M., Santhanam, R., Kwok, R. C.-W., & Yue, W. T. (2022). Could gamification designs enhance online learning through personalization? Lessons from a field experiment. *Information Systems Research*, 1–23. https://doi.org/10.1287/isre.2022.1123

Liao, Y., Vitak, J., Kumar, P., Zimmer, M., & Kritikos, K. (2019). Understanding the role of privacy and trust in intelligent personal assistant adoption. In N. G. Taylor, C. Christian-Lamb, M. H. Martin, & B. Nardi (Eds.), *Information in contemporary society* (Vol. 11420, pp. 102–113). Springer International Publishing. https://doi.org/10.1007/978-3-030-15742-5_9

Liu, D., Santhanam, R., & Webster, J. (2017). Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS Quarterly*, *41*(4), 1011–1034. https://doi.org/10.25300/MISQ/2017/41.4.01

Looyestyn, J., Kernot, J., Boshoff, K., Ryan, J., Edney, S., & Maher, C. (2017). Does gamification increase engagement with online programs? A systematic review. *PLOS One*, *12*(3), e0173403. https://doi.org/10.1371/journal.pone.0173403

Lopatovska, I., Griffin, A. L., Gallagher, K., Ballingall, C., Rock, C., & Velazquez, M. (2020). User recommendations for intelligent personal assistants. *Journal of Librarianship and Information Science*, *52*(2), 577–591. https://doi.org/10.1177/0961000619841107

Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286–5297). https://doi.org/10.1145/2858036.2858288

Maeda, J. (n.d.). *Frequency level checker*. Retrieved January 16, 2022, from https://chuta.cegloc.tsukuba.ac.jp/flc/index.html

Majuri, J., Koivisto, J., Hamari, J. (2018). Gamification of education and learning: A review of empirical literature. In *Proceedings of the 2nd International GamiFIN Conference, GamiFIN 2018* (pp. 11–19).

Marczewski, A. (2018). *Even Ninja monkeys like to play: Gemification, game thinking and motivational design* (Unicorn ed.).

Mazarakis, A. (2015). Using gamification for technology enhanced learning: The case of feedback mechanisms. *Bulletin of the IEEE Technical Committee on Learning Technology*, *4*(17), 6–9. https://tc.computer.org/tclt/wp-content/uploads/sites/5/2018/01/2_Mazarakis.pdf

Mazarakis, A. (2021). Gamification reloaded: Current and future trends in gamification science. *i-com*, *20*(3), 279–294. https://doi.org/10.1515/icom-2021-0025

Mazarakis, A., & Bräuer, P. (2018). Gamification is working, but which one exactly? Results from an experiment with four game design elements. In *Proceedings of the Technology, Mind, and Society* (Vol. 22, p. 1). https://doi.org/10.1145/3183654.3183667

Mazarakis, A., & Bräuer, P. (2022). Gamification is working, but which one exactly? Results from an experiment with four game design elements. *International Journal of Human–Computer Interaction*, 1–16. https://doi.org/10.1080/10447318.2022.2041909

McTear, M., Callejas, Z., & Griol, D. (2016). *The conversational interface*. Springer International Publishing. https://doi.org/10.1007/978-3-319-32967-3

Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, *71*, 525–534. https://doi.org/10.1016/j.chb.2015.08.048

Monterrat, B., Desmarais, M., Lavoué, É., & George, S. (2015). A player model for adaptive gamification in learning environments. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial intelligence in education* (Vol. 9112, pp. 297–306). Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_30

Moussalli, S., & Cardoso, W. (2020). Intelligent personal assistants: Can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning*, *33*(8), 865–890. https://doi.org/10.1080/09588221.2019.1595664

Murad, C., Munteanu, C., Clark, L., & Cowan, B. R. (2018). Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human–Computer Interaction with Mobile Devices and Services Adjunct – MobileHCI'18* (pp. 269–276). https://doi.org/10.1145/3236112.3236149

Namaghi, S. A. O., Safaee, S. E., & Sobhanifar, A. (2015). The effect of shyness on English speaking scores of Iranian EFL learners. *Journal of Literature, Language and Linguistics*, *12*, 22–28.

Nobles, A. L., Leas, E. C., Caputi, T. L., Zhu, S.-H., Strathdee, S. A., & Ayers, J. W. (2020). Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *NPJ Digital Medicine*, *3*(1), 11. https://doi.org/10.1038/s41746-019-0215-9

Ortiz-Rojas, M., Chiluiza, K., & Valcke, M. (2019). Gamification through leaderboards: An empirical study in engineering education. *Computer Applications in Engineering Education*, *27*(4), 777–788. https://doi.org/10.1002/cae.12116

Planner, A. (2018, July 5). *How many users should you expect for your voice skill?* https://www.screenmedia.co.uk/news/how-many-users-should-you-expect-for-your-voice-skill/

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Pyae, A., & Scifleet, P. (2018). Investigating differences between native English and non-native English speakers in interacting with a voice

user interface: A case of Google Home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (pp. 548–553). https://doi.org/10.1145/3292147.3292236

Pyae, A., & Scifleet, P. (2019). Investigating the role of user's English language proficiency in using a voice user interface: A case of Google Home smart speaker. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). https://doi.org/10.1145/3290607.3313038

Robson, K., Plangger, K., Kietzmann, J. H., McCarthy, I., & Pitt, L. (2015). Is it all a game? Understanding the principles of gamification. *Business Horizons*, *58*(4), 411–420. https://doi.org/10.1016/j.bushor.2015.03.006

Ruan, S., Jiang, L., Xu, J., Tham, B. J.-K., Qiu, Z., Zhu, Y., Murnane, E. L., Brunskill, E., & Landay, J. A. (2019). QuizBot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3290605.3300587

Ruan, S., Jiang, L., Xu, Q., Liu, Z., Davis, G. M., Brunskill, E., & Landay, J. A. (2021). EnglishBot: An AI-powered conversational system for second language learning. In *26th International Conference on Intelligent User Interfaces* (pp. 434–444). https://doi.org/10.1145/3397481.3450648

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Publications.

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Sailer, M., & Sailer, M. (2021). Gamification of in-class activities in flipped classroom lectures. *British Journal of Educational Technology*, *52*(1), 75–90. https://doi.org/10.1111/bjet.12948

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380. https://doi.org/10.1016/j.chb.2016.12.033

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human–Computer Studies*, *74*, 14–31. https://doi.org/10.1016/j.ijhcs.2014.09.006

Skidmore, L., & Moore, R. K. (2019). Using Alexa for flashcard-based learning. In *Interspeech 2019* (pp. 1846–1850). https://doi.org/10.21437/Interspeech.2019-2893

Speech Synthesis Markup Language (SSML) Reference | Alexa Skills Kit (n.d.). Amazon (Alexa). Retrieved July 17, 2021, from https://developer.amazon.com/en-US/docs/alexa/custom-skills/speech-synthesis-markup-language-ssml-reference.html

Tan, M., & Hew, K. F. (2016). Incorporating meaningful gamification in a blended learning research methods class: Examining student learning, engagement, and affective outcomes. *Australasian Journal of Educational Technology*, *32*(5), 19–34. https://doi.org/10.14742/ajet.2232

Tejedor-Garcia, C., Escudero-Mancebo, D., Cardenoso-Payo, V., & Gonzalez-Ferreras, C. (2020). Using challenges to enhance a learning game for pronunciation training of English as a second language. *IEEE Access*, *8*, 74250–74266. https://doi.org/10.1109/ACCESS.2020.2988406

Tejedor-Garcia, C., Escudero-Mancebo, D., Gonzalez-Ferreras, C., Cámara-Arenas, E., & Cardenoso-Payo, V. (2016). Improving L2 production with a gamified computer-assisted pronunciation training tool, Tiptoptalk! In *Proceedings of IberSPEECH* (pp. 177–186).

Thiebes, S., Lins, S., & Basten, D. (2014). Gamifying information systems—A synthesis of gamification mechanics and dynamics. In *ECIS 2014 Proceedings* (pp. 1–17).

Tyack, A., & Mekler, E. D. (2020). Self-determination theory in HCI games research: Current uses and open questions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–22). https://doi.org/10.1145/3313831.3376723

Vansteenkiste, M., & Ryan, R. M. (2013). On psychological growth and vulnerability: Basic psychological need satisfaction and need frustration as a unifying principle. *Journal of Psychotherapy Integration*, *23*(3), 263–280. https://doi.org/10.1037/a0032359

Vitkauskaitė, E., & Gatautis, R. (2018). Points for posts and badges to brand advocates: The role of gamification in consumer brand engagement. In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2018.143

Werbach, K., & Hunter, D. (2012). *For the win: How game thinking can revolutionize your business*. Wharton Digital Press.

Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P. R., Clark, L., & Cowan, B. R. (2020). See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers. In *22nd International Conference on Human–Computer Interaction with Mobile Devices and Services* (pp. 1–9). https://doi.org/10.1145/3379503.3403563

Young, H. (2014, November 7). Do young people care about learning foreign languages? *The Guardian*. https://www.theguardian.com/education/2014/nov/07/-sp-do-young-people-care-about-learning-foreign-languages-data

Zellou, G., Cohn, M., & Ferenc Segedin, B. (2021). Age- and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication*, *5*, 600361. https://doi.org/10.3389/fcomm.2020.600361

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

About the authors

Paula Bräuer is a research assistant and PhD candidate at Kiel University since 2018. She has a master's degree in business informatics and researches gamification in various contexts, such as augmented reality and intelligent virtual assistants.

Athanasios Mazarakis is a post-doc for Web Science at ZBW–Leibniz Information Centre for Economics and has been working on gamification and incentives in the interdisciplinary field between computer science, economics, and psychology for over a decade. Numerous publications on gamification and successful workshop organizations complete his competence profile.