Research Article / Araştırma Makalesi

INVESTIGATION OF THE USABILITY AND USER EXPERIENCE OF MOBILE LANGUAGE LEARNING APPLICATIONS: BUSUU, DUOLINGO, AND MEMRISE

MOBİL DİL ÖĞRENME UYGULAMALARINDA KULLANILABİLİRLİK VE KULLANICI DENEYİMİNİN İNCELENMESİ: BUSUU, DUOLINGO, MEMRISE

Ozan SOYUPAK[1]
Haluk İPEK[2]

ORCID: O.S. 0000-0003-4316-1712
H.İ. 0009-0005-1020-7990

Corresponding author/Sorumlu yazar:
[1] Ozan Soyupak
Marmara University, Türkiye
E-mail/E-posta: ozansoyupak@gmail.com

[2] Haluk İpek
Marmara University, Türkiye
E-mail/E-posta: halukipek16@gmail.com

Abstract
This study explores the effects of usability and user experience of mobile language learning applications on the learning process by evaluating three popular applications Busuu, Duolingo, and Memrise. The data were collected by user interviews and the diary method. With the help of the think-aloud method, user interviews were actualized, and in addition to that the researcher kept a diary by trying all three applications for 10 days. Nielsen's Heuristics and an adapted version of these heuristics were used within the framework of this study. The researcher's diaries were coded with the help of heuristics. Findings emphasize the strong connection between the user experience and mobile language learning applications' efficiency. Results indicate that even the adapted version of heuristics has shortages, and it is needed to be developed for future researches. As a result, investigation of user experience is recommended to be evaluated separately for the different phases of application usage in further studies.

Keywords: Mobile Learning, User experience, Usability, MALL, Heuristics

Öz
Bu çalışma, mobil dil öğrenme uygulamalarının kullanılabilirlik ve kullanıcı deneyiminin öğrenme süreci üzerindeki etkilerini üç popüler uygulama olan Busuu, Duolingo ve Memrise'ı değerlendirerek araştırmaktadır. Veriler kullanıcı görüşmeleri ve günlük yöntemi ile toplanmıştır. Sesli düşünme yöntemi yardımıyla kullanıcı görüşmeleri gerçekleştirilmiş ve buna ek olarak araştırmacı her üç uygulamayı da 10 gün boyunca deneyerek günlük tutmuştur. Bu çalışma çerçevesinde Nielsen'in Sezgisel Yöntemleri ve bu sezgisel yöntemlerin uyarlanmış bir versiyonu kullanılmıştır. Araştırmacının günlükleri sezgisel yöntemlerin yardımıyla kodlanmıştır. Bulgular, kullanıcı deneyimi ile mobil dil öğrenme uygulamalarının verimliliği arasındaki güçlü bağlantıyı vurgulamaktadır. Sonuçlar, sezgisel yöntemlerin uyarlanmış versiyonunun bile eksiklikleri olduğunu ve gelecekteki araştırmalar için geliştirilmesi gerektiğini göstermektedir. Sonuç olarak, ileriki çalışmalarda kullanıcı deneyiminin farklı uygulama kullanım aşamaları için ayrı ayrı değerlendirilmesi önerilmektedir.

Anahtar Kelimeler: Mobil Öğrenme, Kullanıcı Deneyimi, Kullanılabilirlik, MALL, Sezgiseller.

INTRODUCTION

Digitalization and developing technology have affected language learning processes as well as every field. First, computers began to take part in the language learning process, and the concept of Computer Assisted Language Learning (CALL) emerged. It dates back to the 1960s. With the developments in online connections and web technologies, the concept of Online Language Learning (OLL) emerged. The evolution continued and the concept of Mobile Assisted Language Learning (MALL) emerged, with mobile devices becoming more involved in daily life. With these developments, the language education and learning process has started to turn into a more complex structure that can be affected by different stakeholders. We can easily say that concepts such as usability, user experience, and user interface have become components of this complex structure and play an important role in its success. The main priority for the success of online learning is the usability of Systems For Electronic learning (SEL) for users and satisfaction degree with the interface (Vlasenko et. al., 2022). The usability of applications created for mobile devices or other platforms, and the holistic experience of the user in this process are among the issues waiting to be designed and researched. As design researchers and designers, we found it valuable to investigate this change and development process of language learning practices.

Traditional to Digital

In general, digitalization and the inclusion of technology in processes are perceived as positive. It may not be right to see digitalization and the integration of technology into the education process as the definitive solution to all problems. There will certainly be some topics that it will help and facilitate, but on the other hand, it should also be appropriate to the nature of education. Díaz Arévalo (2022) stated that although technological developments and language learning applications have become widespread, teacher-student interaction is essential in the learning process, and applications have functions to introduce words and content. The level of technology and complexity should not intimidate any stakeholder who will use this tool. According to Carraher Wolverton & Guidry Hollier's (2019) survey of university lecturers, less technology should be used to simplify learning, and online courses should be simpler. Teachers as well as students are important stakeholders in these systems and their views should be considered. In addition, these stakeholders may not have a homogeneous distribution and the same opinion, there may be sub-user groups with different views. In Ironsi's (2022) study with language instructors and pre-service teachers on the use of Google Meet as a simultaneous language learning tool, pre-service teachers stated that they did not find it efficient and effective, while language instructors stated the opposite.

After digital tools are included in the language learning processes, comparing the old method with the new one, transferring the old to the new, or transforming it are among the perspectives encountered. In this context, studies comparing the effectiveness of digital language learning tools with the traditional language learning process are carried out. One of the most used language learning mobile applications is Duolingo. Rachels & Rockinson-Szapkiw (2018) compared students learning Spanish through face-to-face instruction and Duolingo in their study, and found no significant difference in grammar and vocabulary of the two groups, recommending Duolingo as a cost-effective option. In some studies, the existing methods in the traditional language learning process have been digitized with the help of applications. Mirzaei et al. (2018) focused on the new vocabulary learning process, digitized and tested the keyword method, which is a vocabulary learning strategy, as a tablet-based application. These comparisons and transfers will help us make sense of mobile-centered education and learning technologies.

MALL and GALL

Mobile devices, which are at the center of technology use in daily life, and language learning applications that are developed especially for these devices are the focus of this research. Mobile Assisted Language Learning (MALL), a subcategory of mobile learning, focuses on the language learning process with the help of mobile devices, regardless of the user's location (Shortt et al., 2021). Burston (2014), in his study examining the development process of MALL, stated that although MALL made significant progress, it lagged behind the developments in learning theory and teaching methodologies and that behaviorist and teacher-centered approaches were dominant in MALL practices.

In addition to this backwardness, the new experience of the MALL has also confronted users with many different challenges.

A clear description of what to do and how to do it in these mobile learning applications, and navigation within the application are among the most common problems encountered. Ishaq et al. (2020) investigated the usability and user experience of a mobile language learning app that was specifically designed for the educational context by Pakistani government and they also underlined the interaction and navigational issues for the younger age group and its loose link with the official educational content. Al-Sabbagh et al. (2019) evaluated the usability of three mobile language learning apps with a sample of Arabic-speaking immigrants who had recently arrived in Sweden and found unclear prescription and navigation problems to be among the most common challenges. Another navigation-related issue was that users restarted the application when they did not understand how to navigate. One of the innovative features provided by these applications is enhancing the users' interaction level, allowing users to interact with each other. However, in Falk et al. (2016) study on Duolingo's interaction feature with 212 participants, although automatic feedback and interaction with the application were found positive and beneficial, just very few users used the interaction with other users feature. As in this case, every new and different feature may not be adopted by the user, so offering experience and comprehending the user's needs are important. In addition to these studies, in their systematic literature study, Ahmad Faudzi et al. (2023) define visual representation, user control, navigation, and signaling as four critical features that in the case of misuse may cause cognitive load to the user.

There may be instances where the user's needs differ even in terms of language learning. To continue their lives, people may have to learn a language even when they have limited sources. Within this context, a number of applications targeting immigrants in Germany have been developed, with minimum technical requirements and compatibility with old phones, German language education, and tips on living in Germany are among the main features of these applications (Godwin-Jones, 2017). In addition to burdening these applications with vital responsibilities such as maintaining life in another culture and country, users should enjoy using these applications and have the desire to use them again. One of the ways to achieve this is gamification. According to Kapp (2012), gamification uses game-based mechanics, aesthetics, and game thinking to engage people, motivate action, promote learning, and solve problems. The inclusion of the concept of gamification in language learning applications has also produced positive results. Gamification Assisted Language Learning (GALL) emerged with CALL's collaboration with the concept of gamification, and Udjaja (2018) revealed in his study that GALL contributed between 20% and 100% to the Japanese learning process. Loewen et al. (2019), in their study in which they examined the Turkish learning process of nine participants during a semester with the Duolingo application, found that there was a positive relationship between the time spent in Duolingo and learning outcomes, and features such as flexibility and gamification were seen as positive. On the other hand, the fact that the developed learning applications do not contain sufficient gamification may cause the users not to complete the learning process (Tundjungsari, 2020).

Usability and User Experience
In this section, first, the concepts of usability and user experience will be shared. With the help of these concepts, previous studies on the design of mobile and learning environments will be revealed. The most common definition of usability is; the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments (International Organization for Standardization, 2018). In usability tests, where real users are asked to perform a genuine task, the efficiency of the interaction and user satisfaction are interpreted (Hamel, 2012). For these evaluations and tests, Heuristics developed by Nielsen have a pioneering feature.

Nielsen's 10 Heuristics is a systematic review of a user interface design for usability, aiming to find usability problems in a user interface design so that they can be addressed as part of an iterative design process (Nielsen, 1994). Heuristic evaluation is an informal method of usability analysis in which a set of evaluators is presented with an interface design and asked to comment on it, and datasets from several evaluators can yield quite good results (Nielsen & Molich, 1990). Rao Naidu et al. (2020) evaluated four

tools designed for the language learning process, namely FluentU, Duolingo, Livemocha, and Hello English, using Nielsen's 10 Heuristics. They evaluated each tool as one of the *"met" "partially met"* and *"not met"* items according to each heuristic item. However, the data obtained as a result of this evaluation is limited and may not result in a richness that will feed the design process because it is over-categorized and simplified. Ishaq et al. (2021) made a low-fidelity prototype of the game-based language learning application they developed and evaluated it using heuristic and think-aloud methods. Four teachers and one game designer participated in this evaluation process. The questionnaire used in the assessment consists of 47 elements and there are 'Yes', 'No', or 'Not sure' options for each element. Although the think-aloud method is one of the methods they use, the evaluators' expertise level in design and user experience, low fidelity prototype usage, and limited answer options in the questionnaire makes it hard to reach the data richness.

Usability evaluations are often confused with user experience, which also encompasses user thoughts, feelings, and preferences (Weichbroth, 2020). User experience, a kind of experience, is the sum of the effects that the user feels before, during, and after interacting with a product or system (Hartson & Pyla, 2018). Hartson & Pyla (2018) have expressed the basic features of the user experience in items as follows.

(1) It is a result of direct and indirect Interaction.
(2) It is related to the totality of effects.
(3) It is felt internally by the user.
(4) It comprises usage context and ecology

Usability aims for effectiveness and efficiency, focused on user behavior and interaction; user experience, on the other hand, aims to meet psychological needs and focuses on experience (Hassenzahl et al., 2021). Hermawati and Lawson (2016) reviewed 70 studies on the customization of Design Heuristics for specific areas and stated that although general Heuristics are suitable for most interfaces and situations, some areas require implementation. They said the lack of validation effort and validation method for this domain-specific Heuristics as an urgent problem.

Apart from the studies that directly use Nielsen's Heuristics, within the perspective of this study, the studies carried out by adapting the heuristics on similar subjects to this study have been examined. Zaharias and Poylymenakou (2009) have developed a questionnaire-based usability evaluation method for e-learning applications, which focuses not only on cognitive but also on affective considerations. The framework developed by Kumar et al. (2020) for the heuristic evaluation of mobile learning applications consists of three stages: planning, conducting, and reporting. In the planning phase; the content of the examination is defined, participants are chosen, hardware and software are chosen, and documents and materials are prepared. In the conducting phase heuristic evaluation is directed, and in the reporting phase the results are put in and possible solutions are identified. Jamaldeen et al. (2018) summarized the current design principles in the mobile learning content, which aimed to create a design guideline specific to mobile language learning applications for blended learning environments, as follows: Design the interface to compensate for the small screen, Navigational strategies, Consider a variety of elements, Spontaneous access of materials, Support multiple device types, Keep the information organized, Multiple pedagogy/activities, Design for personalized learning, Incorporate collaborative activities. While applying good design guidelines alone is a good start, it is not a substitute for system usability evaluation, usability depends on the context and is shaped by the interaction between users, tasks, and system intent (Granić and Ćukušić, 2011).

Kumar and Goundar (2019) reconsidered and developed Nielsen's 10 heuristics with a mobile learning application perspective, and also proposed 3 new heuristics: Selection driven commands, content organization, and visual representation. In this study, which used heuristics as one of the evaluation tools, it was tried to find answers to the following questions.

(1) What kind of user experience do the most used mobile language learning apps offer? What are the

problematic aspects? What are the successfully resolved features?
(2) What are the components of the framework that will be used to design a mobile application that will facilitate the language learning process?
In this framework, this study approaches the language learning process holistically and focuses on the user experiences of groups of young adults at different proficiency levels.

METHOD

Design research projects can be generalized into two phases as generative research and evaluative research. Instead of this two-phase model, the first phase of which is defined as formative, discovery, exploratory, or generative, and the second phase consists of iterative testing, or evaluation research, Hanington (2007) proposed a three-phase model. These three phases are explore, generate and evaluate, which overlap each other. Exploratory research is carried out in the early stages of the design process, especially for designers to gain knowledge in an area they are unfamiliar with (Martin and Hanington, 2012). It includes ethnographic methods and other design methods such as diary studies, observation, and surveys.

In this study, which can be called exploratory research; it was tried to find answers to the research questions by examining three mobile language learning applications. These apps are Duolingo, Memrise, and Busuu. The reason for choosing these three applications is that they are the 3 applications most commonly used in the potential sample in the unstructured pre-interviews. Different applications may be preferred more in different samples and in different geographies. In these cases, the price of the application, the use and advice of the close environment may be among the reasons for the preference. Data were collected with two different tools, user interviews, and researcher diaries. The participants of the study consisted of 3 female, 7 male, aged between 20-24 years with language proficiency levels ranging from pre-intermediate to upper intermediate (Table 1).

Table 1. Participants' Data.

	Age	Gender	Proficiency Level
U1	22	Female	Pre-intermediate
U2	21	Male	Intermediate
U3	23	Male	Upper Intermediate
U4	20	Male	Pre-intermediate
U5	21	Male	Pre-intermediate
U6	24	Male	Intermediate
U7	20	Female	Intermediate
U8	22	Male	Upper Intermediate
U9	20	Female	Pre-intermediate
U10	21	Male	Pre-intermediate

User Interviews

In this data collection process, semi-structured interviews were conducted with 10 participants who were priorly informed about the research and gave their consent to use the relevant data. In the first stage of the interviews, the participants' backgrounds, approaches to language learning and level of familiarity with technology were collected. Each user was interviewed individually. *"Think-aloud"* method was used after the semi-structured interviews. *"Think-aloud"* is a detailed observation method used to

examine mental processes in which participants are asked to make verbal comments while working on a task (Lewis, 1982). *"Think-aloud"* is a method used when users test an application or design proposal. Cotton and Gresty (2006), in their study evaluating the use of the think-aloud method in the field of e-learning, stated that although the analysis process might be difficult and time-consuming, the data they obtained with this method were information that researchers could overlook with other methods, and they emphasized that this method yielded extremely rich and explanatory data. There is no need to have a finished product to test with and learn from the user, existing products, early designs, general interaction concepts, detailed interactions, and versions of your products can also be tested at any stage of the process (Perea and Giner, 2017). Since this study is mainly exploratory research and does not include a design proposal yet, *"the think aloud"* method was used when using the three applications mentioned above. The snowball sampling method was used, and interviews were conducted with people who actively use at least one of these three applications. In the think aloud phase, each user was asked to use the application and while the users were experiencing the application, the researcher took audio recordings of the users' experiences, took notes on the users' expressions, positive and negative features, and recorded relevant screenshots.

Diaries and Logs
Rose (2019) defines the differences between *"diaries"*, *"logs"* and *"journals"* from a language learning research perspective. According to Rose (2019) diaries have an open-ended, unstructured, and reflective nature while logs are concentrated on collecting the more specific information needed and are therefore more structured and constrained.

The biggest advantage of the diary method is that it obtains longitudinal information in the natural context of the interaction (Lallemand, 2012).

A pro account was created for all 3 applications by the first researcher and used for a week to become familiar with the applications. Afterward, the first researcher took notes by using all three applications for at least half an hour every day for 10 days. He took these notes in two different ways and within the perspective of two different identities.

On odd days, when he saw a problem on the screens, interface, or interaction, he immediately took a screenshot. He has noted problems alongside and above these screenshots. Expert reviews are the evaluation of interfaces conducted by experts in the field or user interface design, using reference points such as design heuristics (Shneiderman et al., 2018). He existed in these odd days, with the identities of expert and evaluator.

On even days, the researcher used each application for half an hour. While using it, he took small notes if he needed it. At the end of the half-hour session in which he used the application, he took note of his experience and reflections. He expressed his own learning experience explicitly. Here, we can say that the researcher exists mostly with a user identity.

Notes taken by the researcher on odd days can be expressed as logs because they are more structured and simultaneous. The more unstructured notes he took after the experience on even days can be expressed as diaries (Table 2).

Table 2. 10 Days Long Diary Process of the Researcher.

	Role	Duration	Actions	Qualification
Odd Days **(Day 1, 3, 5, 7, 9)**	Expert Evaluator	+90 minutes (at least 30 minutes for each app)	Expert examination and screenshot taking when there is a problem with interface or interaction	Log input

Even Days (Day 2, 4, 6, 8, 10)	User	+90 minutes (at least 30 minutes for each app)	Taking reflective notes about user experience	Diary input

FINDINGS

User Interviews

A total of 10 people were interviewed, including 4 Busuu users, 3 Memrise users, and 3 Duolingo users. The interviews were evaluated on an application-specific basis. First of all, audio recordings of all interviews were translated into text. In the analysis process, MAXQDA, a qualitative data analysis program, was used. On MAXQDA, the interview records were coded with an open coding method. Open coding is accomplished by dividing the data into meaningful expressions and defining them. The goal is to develop rich codes to describe data (Vollstedt and Rezat, 2019). The themes that stand out in the data collected from the users and important quotations were transferred to Miro. Miro is a kind of online whiteboard application. Axial coding is a qualitative research method that involves associating data and establishing connections between data (Simmons, 2018). While we were trying to create the relationship between these data, we noticed that there was also a time-wise relationship between them. We noticed that the data is distributed over three main time sets: application usage time, before and after. Application usage time, on the other hand, was divided into components as study time, before and after. We performed time-wise axial coding of us with the help of these.

Visual summaries of the analysis process are given in the figures below. We noted the positive feedback of the users about the application on the green post-its, the negative feedback on the red post-its, the direct quotations on the blue post-its, and the wishes/requests shared by the users on the yellow posts.

Figure 1 contains a visual summary of the outputs of Busuu's analysis studies. If we start from the registration process, the questions asked at this stage, the time, and language options provide a special initial configuration for the user. In addition, the placement test performed at this stage and starting the lesson directly after it allows the user to enter the application in a smooth, target-oriented manner.

The difference between the visual identities of the screens onboarding and immediately after was expressed by a user as follows.

There is something about this screen, I felt like I switched to another application because the design of this page is very different from the pages I encountered at the beginning.

Users gave negative feedback about the paid membership promotions they encountered as soon as they started using the application. The absence of the dark theme option was also noted by many users as an important shortcoming.

In the lecture phase, the categorization of the content as words and grammar was seen positively by the users. In addition, sharing additional information about the culture of the society is stated as a feature that differentiates Busuu. Supporting the content with visuals and texts, different question types, multi-faceted feedback, and the progress bar that gives feedback on the stage of the process are other positive features emphasized about the lecture phase.

Other characteristic features are the ability to socialize, study with friends, and help others. Helping others, in particular, is an innovative feature not included in other apps, where users correct other users' exercises that are in their native language or languages they know very well. The possibility of being exposed to frequent notifications when the application is not used was emphasized as a negative situation by the users.

You can find the visual summary of the outputs of Memrise's analysis below (Figure 2). Similar to Busuu, the problem of promoting paid memberships too often, in the beginning, was encountered here as well. A user stated that after achieving certain successes in the application and after using sessions of the application, a paid membership promotion would be more successful.
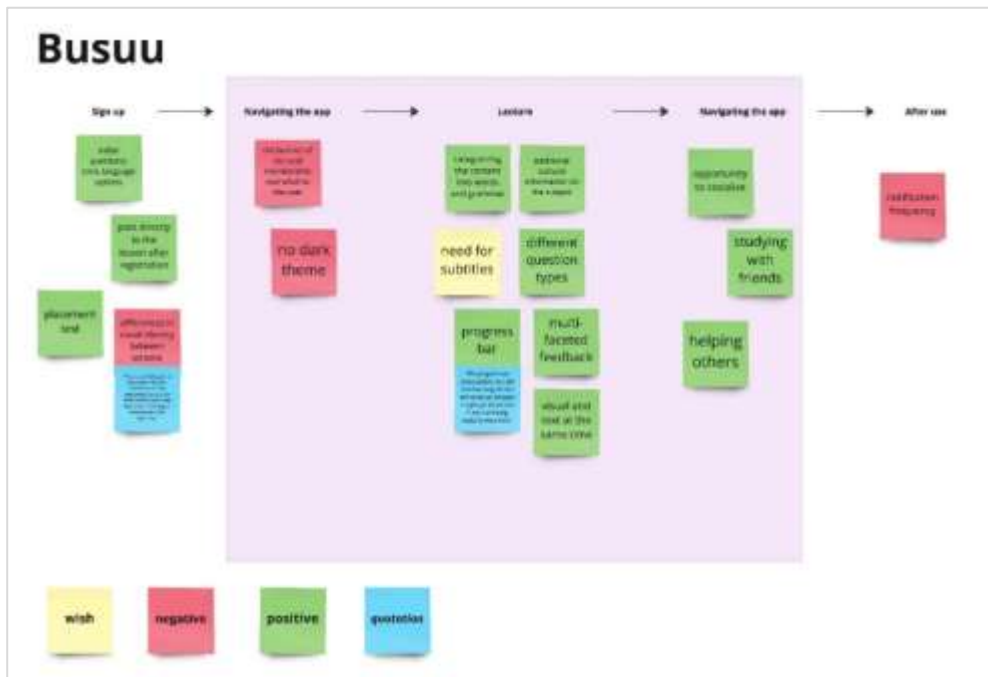

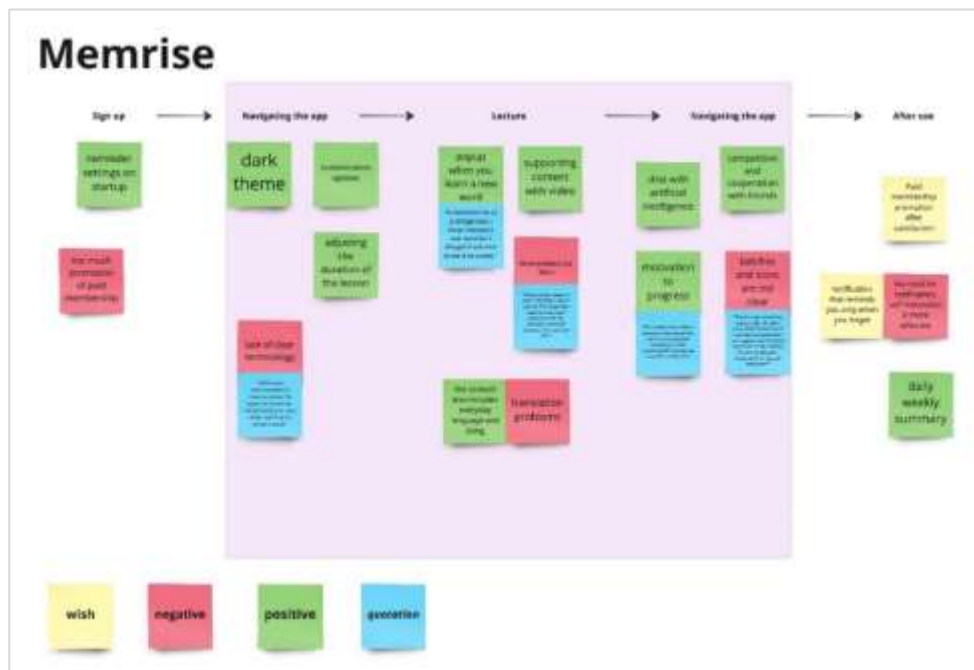**Figure 1.** Busuu analysis process outputs visual summary.

Figure 2. Memrise analysis process outputs visual summary.

Some expressions and design elements are not fully understood. A user expressed the lack of clarity in terminology as follows.

847

"I didn't quite understand what it meant to master the word. So, how do we classify mastery or what do we need to do to master a word?"

Similarly, there are points in batches and icons that users cannot understand.

"There is a sign on the first lesson, it says 14. I don't know what it is, but it could have been expressed more meaningfully when I finished the first lesson. For example, the color of that green button could change. Still looks green."

Another item that other users had difficulty in making sense of was the points they earned during the lesson.

"Points, exactly. I wouldn't have noticed if you hadn't told me. They could have made the score more understandable, for example, in terms of estimation. I've never seen him."

All these unclear, incomprehensible, and hard-to-understand points cause interruptions in the user experience. The options provided by adjusting the duration of the lesson and many other features allow the user to customize the experience according to themselves are considered successful by users. Supporting the content with videos and including everyday language and slang are other features of

Memrise course content emphasized by users. On the other hand, some translation problems in the course content were reported as incompleteness by the users. Outside the classroom, chatting with AI, competition, and cooperation with friends, daily weekly summaries are among the memrise features that users are happy with.

Below is a visual summary of Duolingo's analysis study (Figure 3). Similar to other applications, the promotion of the paid membership and the discomfort of the users were encountered here as well. In fact, a user defined the constant discount of a paid membership on the main screen as a pressure to buy.

While users found the home screen and learning process to be defined as a path useful for following their processes, they stated their discomfort about the locked content and lessons on the path. One user stated that there are too many game items.

The fact that the learning process includes so many game elements were not considered appropriate by the user. Similarly, other users have stated that they are uncomfortable with the excess of elements such as badges and strikes. The excess of these elements also causes complexity and reduces comprehensibility.

Different users have stated that the tabs at the bottom of the page that provide navigation with other pages are excessive and not understandable as follows.

"Even without 6 tabs, these could have been done by minimizing. All social things could be together. These "Daily Quests" could be clicked to go to the "Feed" section or from the profile, such as notifications, for example, this part. It's getting very crowded. And because they're all colorful..."

"There is a birdhouse, it takes you to the home page, there is no problem with that. Broken heart picture It tells about the mistakes I made, but I couldn't detect them without clicking it. The shield shows my ranking. One female picture shows my own profile. I don't think any of this is understandable."

We encountered that the progress bar, in other words, knowing the time it takes to complete the exercise motivates users to complete it. In addition, users stated that they care about numerical data which is shared after the exercise such as how many words were learned and how long they spent in the exercise.

Interaction with friends, emphasizing that the activities carried out with friends are more fun,

congratulating success, and motivation to beat someone or to compete with each other are among the data we have reached about the social dimension of the application.
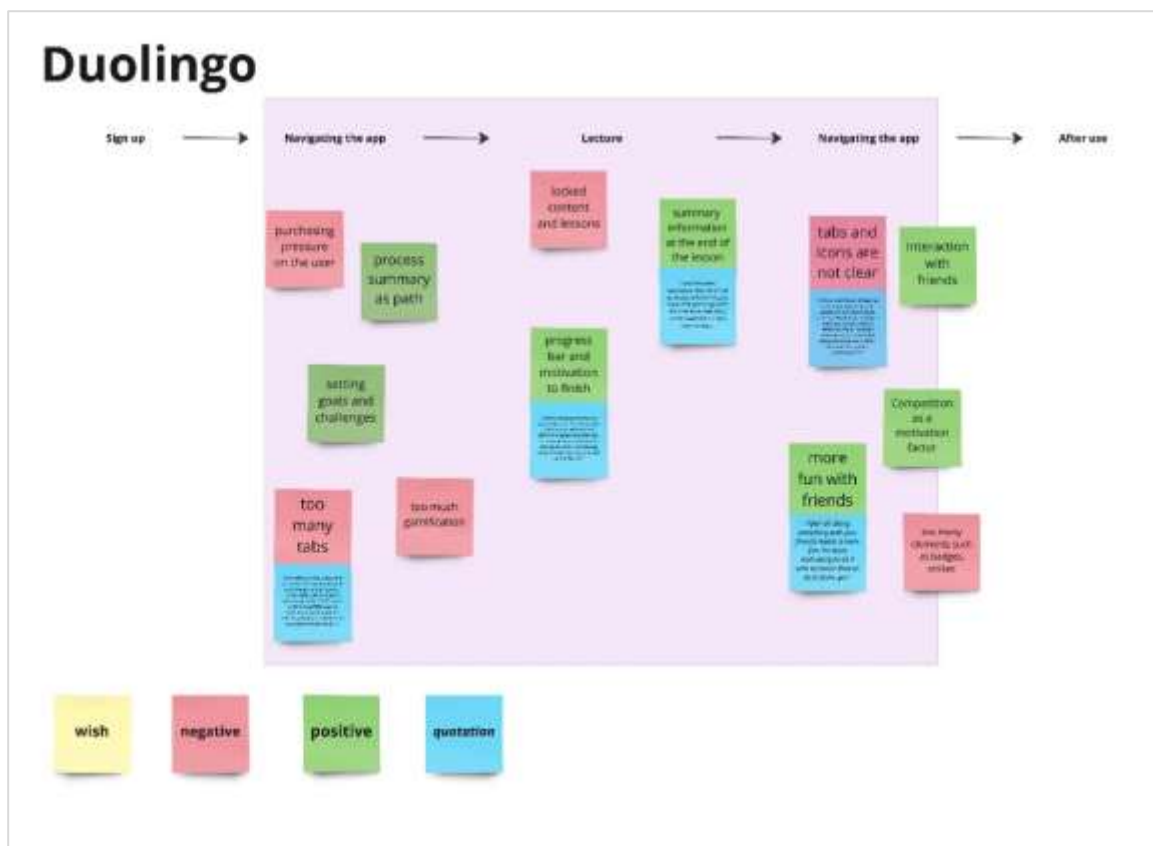

Figure 3. Duolingo analysis process outputs visual summary.

The weaknesses and strengths of the three applications imply the positive and negative attributes of language learning apps in terms of user experience. Unclear badges and icons' usage, frequent notifications, paid membership promotions were evaluated negatively by the users. On the other hand, being able to follow personal progress, socialization and collective learning with friends, customization for learning content, dark themed screens were apricated by the users.

Table 3. Detailed User Experience Comparison of Busuu, Memrise and Duolingo

	Weaknesses	Strengths
Busuu	Paid membership promotions	Initial questions for customization
	Absence of dark theme option	Initial placement test
	Visual identity differences in between screens	Direct pass to lessons after registration
	Frequent notifications	Grammar and word content categorization option
	Lack of subtitles	Progress bar
		Different question types
		Multi-faceted feedback
		Visual and text combination

		Socialization opportunity
		Collective learning with friends
Memrise	Paid membership promotions	Dark theme option
	Unclear terminology usage	Initial reminder settings
	Translation problems	Customization option
	Incomprehensive items	Being able to adjust course duration (self-phased)
	Unclear badges and icons	Word pop-up
	Notifications	Everyday language content
		Video support
		AI chat opportunity
		Collective learning with friends
		Daily weekly summary
Duolingo	Purchasing pressure	Process summary
	Over gamification	Progress bar
	Too many tabs	Customizable goals
	Locked content	After-course summary
	Unclear tabs and icons	Collective learning with friends (competitions, interactions, challenges)
	Too many visual elements	

Diaries and Logs

All three applications were used by the first researcher each day for a total of 10 days. Each session lasted around thirty minutes. The data analyzed in this section consists of notes taken after 30 sessions and 15 hours of use in total. The researcher created the dataset by taking notes as an expert identity (analyze) on odd days and as a user identity (reflection) on even days. We used the evaluation tool obtained by Kumar and Goundar (2019) by revising Nielsen's heuristics. All the notes were transferred to the MAXQDA program, a total of 30 documents. Positive and negative subcodes were also created for each item in the 13-item Koumar and Goundar's evaluation tool. If a successful situation or application related to the relevant evaluation item was noted, it was coded with positive; and if a problem was noted, it was coded with negative.

For all three applications, the notes taken with the expert ID and the number of codes for these notes are more than those taken with the user ID. Out of a total of 153 codings, 98 are in the analysis notes and 55 are in the reflection notes. The ratio is close to 1 to 2. The reason for this may be that one by one and more detailed problem-specific notes are taken at the time of use in the analysis notes, and more general notes are taken in the reflection notes after the application use. 114 of the codes are negative and 39 are positive. We can say that a significant part of the notes taken during the research consists of problem areas. Among the three applications, Memrise (60) was the most coded and Duolingo (41) was the least coded. There is approximately a 1/3 relationship between positive and negative codes for Duolingo and Memrise. In Busuu, the rate of positive codes is more than 1/3.

In the table below, there is a detailed report based on each evaluation item (Table 3). There are striking

differences between the coding numbers of 13 items. "Match between the system and real world", "Recognition rather than recall" and "Selection driven commands" are coded only a few times. "Visibility of the system status", "Visual representation" and "Content organization" were the most coded items. Most of the coding for "Visual representation" is in the analysis notes. "Visibility of the system status" codes are more evenly distributed between analysis and reflection. It is normal for the codes related to visual elements to be mostly among the notes taken with the help of screenshots instantly.

Table 4. Detailed Numerical summary of Diary's work

	Analyse						Reflection						SUM
	Busuu		Duolingo		Memrise		Busuu		Duolingo		Memrise		
	+	-	+	-	+	-	+	-	+	-	+	-	
Aesthetic and minimalist design	0	1	1	0	3	1	0	1	0	1	1	0	9
Consistency and standards	0	3	0	1	0	2	0	0	0	1	0	0	7
Content organization	2	4	0	3	1	5	1	1	1	0	0	6	24
Error prevention	1	2	0	1	1	0	0	1	0	0	1	1	8
Flexibility and efficiency of use	1	2	1	1	0	2	0	1	1	2	0	2	13
Help and documentation	0	0	0	2	1	2	0	0	0	1	0	1	7
Help users recognize diagnose and recover from errors	1	0	2	0	0	2	0	1	0	1	1	2	10
Match between the system and the real world	0	1	0	1	0	0	0	0	0	0	0	0	2
Recognition rather than recall	0	1	0	1	0	2	0	0	0	0	0	0	4
Selection driven commands	0	1	0	0	0	0	0	1	0	0	0	1	3
User control and freedom	1	2	0	1	0	2	0	1	0	2	0	2	11
Visibility of the system status	4	2	3	2	1	1	1	3	1	1	4	1	24
Visual representation	3	5	0	7	0	8	0	3	0	2	0	3	31
SUM	13	24	7	20	7	27	2	13	3	11	7	19	153
	37		27		34		15		14		26		

We think that it is not a problem if an element in the table is coded as few or not at all. For example, for Busuu, there is no coding in the *"help and documentation"* item. Since there was no problem with this during use, any note was not coded as negative. On the other hand, the absence of positive coding does not mean that there is no positive experience. On the contrary, the experience may be fluent without the need for help. Another point that draws our attention is that it is too generalizing to evaluate an application with a single positive or negative statement specific to an evaluation item.

As seen in the table above, an application has both positive and negative codes specific to an evaluation item. While there are problems in *"help and documentation"*, for example, in the stages of creating a new user and starting to use the application, it may have good documentation for the course content and exercises. Or multiple negative codes belonging to the same evaluation item of the same application may belong to different screens and different stages of the usage process.

In addition, we thought that some items were worth coding, but none of the 13 items and their explanations were fully met. For example, *"content organization"* is included in the evaluation tool as an item. The quality of the content, which senses the content appeals to, and what elements it consists of remain outside the definition in the evaluation tool. We think that all these criteria should be included in the evaluation process for every application, especially in a learning and language learning application.

CONCLUSION

In this study, we evaluated the user experience of the most popular mobile language learning apps by focusing on their positive and negative features, and tried to outline specific user experience components for language learning.

Within this scope, we have already shared the definitions and perspectives of usability and user experience. In this study, we adopted more open-ended methods in the evaluation of the interviews, while we adhered to a 13-item heuristic evaluation tool in the diary evaluations. We think it would be beneficial to use more expanded tools that include the richness of both perspectives and to include beyond efficacy and effectiveness in the evaluations.

Mobile devices and mobile technologies have some features and constraints by their nature, and we need to include them in evaluation processes and tools. The design constraints of mobile technology can be specified as limited screen space, battery life, storage, memory and communication capabilities, usage in different physical and social contexts, and hosting different screen technologies and sizes (Benyon, 2019). These properties can change very radically depending on the context, since they are technical values, numerical methods can also be developed to measure and compare them. Despite all these limitations, the user's expectation is not modest. It is expected of the product owners to develop the best suitable solution. This study shows us for language learning apps, it's important to form a consistent visual language in user interface, to provide customization both in user interface and learning content, to present personal progress of the user, to include multi-faceted information transmission, to have a dark mode, to provide a space for learning with friends and to use clear icons. Over-gamification, frequent notifications, inadequate subtitling, complex terminology, unclear visuals and aggressive paid-membership policy negatively affect the user experience.

Even if we use an expanded version of Nielsen's heuristics specific to mobile learning, we think that some titles and important items are still out of scope. We have shared examples of these above. Interaction and quality of interaction, privacy, and similar items can also be added.

The experience consists of different stages, and not every element may be fully covered at each stage. More detailed assessments may be required for each stage of the experience. For these reasons, we think that there is a need for time-wise axial coding and heuristic evaluation, which we use in the analysis phase of user interviews. To facilitate the evaluation process, we think it would be appropriate to clearly identify these potential additives and even group them under categories. We think it would be useful to

have more comprehensive documents, including examples, for more novice evaluators.

Usability is key to the success of e-learning applications, and if there is a problem with the user interface, people will walk away, get frustrated, and stop using it (Ardito et al., 2006). If there is a usability problem in an e-learning application, it will harm the user's learning process, no one wants to spend more time learning how to use the software instead of learning the content (Wong et. al, 2003). For these reasons, even a small error that may occur in this section may be difficult to compensate for. It may be appropriate for the whole team to adopt these concepts and to inform not only the designers and experts but also the developers about these concepts.

This study has potential limitations. The research sample were young adult user group, who were already familiar with the target language and had certain level of technological skill and mobile app usage. User experience of people with different language and technological skills and different aged results may not be the same. Experience of an elder person or a kid, or a total beginner or a professional language learner naturally will be different from the sample. For example, over-gamification might not be a problem for a younger age group, or professional language learners might seek more than everyday language.

Most research in the field of HCI is short-term studies, and on the contrary, the importance of examining long-term use is emphasized by different sources (Karapanos et al., 2021). The researcher's diary part of this study is valuable and differentiated as it examines a 10-day period. We recommend that similar studies that are spread over time and that are specific to different user groups should be carried out and that evaluation tools should be developed and updated with the help of these studies.

Acknowledgement

Data avaliability
All data are available from the corresponding author upon reasonable request.

Disclosure statement
There is no potential conflict of interest associated with this research study.

REFERENCES
Ahmad Faudzi, M. C. (2023). Investigating the user interface design frameworks of current mobile learning applications: A systematic review. *Education Sciences, 13*(1), 94-118.
Al-Sabbagh, K., Bradley, L., & Bartram, L. (2019). Mobile language learning applications for Arabic speaking migrants–a usability perspective. *Language Learning in Higher Education, 9*(1), 71-95.
Ardito, C., Costabile, M. F., Marsico, M. D., Lanzilotti, R., Levialdi, S., Roselli, T., & Rossano, V. (2006). An approach to usability evaluation of e-learning applications. *Universal Access In The Information Society, 4*(3), 270–283. doi:10.1007/s10209-005-0
Benyon, D. (2019). *Designing user experience.* Pearson UK.
Burston, J. (2014). MALL: The pedagogical challenges. *Computer Assisted Language Learning, 27*(4), 344-357.
Carraher Wolverton, C., & Guidry Hollier, B. (2019). A minimalist design for distance learning. *International journal of educational management, 33*(7), 1457-1465.
Cotton, D., & Gresty, K. (2006). Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology, 37*(1), 45-54. doi:10.1111/j.1467-8535.2005.00521.x
Díaz Arévalo, D. (2022). Gamification strategy and its impact in the English learning vocabulary. *Master's thesis*. La Libertad: Universidad Estatal Península de Santa Elena.
Falk, S., Götz, S., Zeyer, T., Stuhlmann, S., & Jones, R. (2016). Interactivity in language learning

applications: a case study based on duolingo. *Interaktivität beim Fremdsprachenlehren und-lernen mit digitalen Medien: Hit oder Hype*, 237-258.

Godwin-Jones, R. (2017). Smartphones and language learning. *Language Learning & Technology, 21*(2), 3-17.

Granić, A., & Ćukušić, M. (2011). Usability testing and expert inspections complemented by educational evaluation: A case study of an e-learning platform. *ournal of Educational Technology & Society, 14*(2), 107-123.

Hamel, M. (2012). Testing aspects of the usability of an online learner dictionary prototype: A product-and process-oriented study. *Computer Assisted Language Learning, 25*(4), 339-365.

Hanington, B. M. (2007). Generative research in design education. *International Association of Societies of Design Research2007: Emerging Trends in Design Research*, (s. 12-15). Hong Kong.

Hartson, R., & Pyla, P. S. (2018). *The UX book: Agile UX design for a quality user experience.* Cambridge, US: Morgan Kaufmann.

Hassenzahl, M., Burmester, M., & Koller, F. (2021). User experience is all there is: twenty years of designing positive experiences and meaningful technology. *i-com, 20*(3), 197-213.

Hermawati, S., & Lawson, G. (2016). Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus. *Applied Ergonomics, 56*, 34-51.

International Organization for Standardization. (2018). Ergonomics of human-system interaction- Part 11: Usability: Definitions and concepts (ISO Standard No. 9241).

Ironsi, C. (2022). Google Meet as a synchronous language learning tool for emergency online distant learning during the COVID-19 pandemic: Perceptions of language instructors and preservice teachers. *Journal of Applied Research in Higher Education, 14*(2), 640-659.

Ishaq, K. R. (2020). Usability and design issues of mobile assisted language learning application. *International Journal of Advanced Computer Science and Applications, 11*(8), 86-94.

Ishaq, K., Rosdi, F., Zin, N. A., & Abid, A. (2021). Heuristic and think aloud method to evaluate the low fidelity prototype of game-based language learning application. *2021 International Conference on Innovative Computing (ICIC)* (s. 1-10). Lahore: IEEE.

Jamaldeen, F. F., Hewagamage, K. P., & Ekanayaka, Y. (2018). Design guidelines for creating mobile language learning applications. *International Journal of Interactive Mobile Technologies, 12*(3), 52-74.

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

Karapanos, E., Gerken, J., Kjeldskov, J., & Skov, M. B. (2021). Introduction to "advances in longitudinal hci research". *Advances in Longitudinal HCI Research* (s. 1-7). içinde Springer. doi:10.1007/978-3-030-67322-2_1

Kumar, B. A., & Goundar, M. S. (2019). Usability heuristics for mobile learning applications. *Education and Information Technologies, 24*, 1819-1833. doi:10.1007/s10639-019-09860-z

Kumar, B. A., Goundar, M. S., & Chand, S. S. (2020). A framework for heuristic evaluation of mobile learning applications. *Education and Information Technologies, 25*, 3189-3204.

Lallemand, C. (2012). Dear diary: Using diaries to study user experience. *User Experience Magazine*. https://uxpamagazine.org/dear-diary-using-diaries-to-study-user-experience/ adresinden alındı

Lewis, C. (1982). *Using the" thinking-aloud" method in cognitive interface design* . Yorktown Heights, NY: IBM TJ Watson Research Center.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: a duolingo case study. *ReCALL, 31*(3), 293-311.

Martin, B., & Hanington, B. (2012). *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design Effective Solutions.* Rockport Publishers .

Mirzaei, S., Wilkinson, B., & Wyra, M. (2018). Usability testing of VLASTA: a vocabulary learning and strategy teaching app. *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, (s. 685-689).

Nielsen, J. (1994). *Usability engineering.* Morgan Kaufmann.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems*, (s. 249-256).

Perea, P., & Giner, P. (2017). *UX design for mobile.* Packt Publishing Ltd.

Rachels, J., & Rockinson-Szapkiw, A. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning, 32*(1-2), 72-89.

Rao Naidu, V., Srinivas, S., Al Raisi, M., & Dattana, V. (2020). Evaluation of hypermedia tools in terms of usability heuristics for English language teaching. *Arab World English Journal (AWEJ) Proceedings of 2nd MEC TESOL Conference*, (s. 133-149).

Rose, H. (2019). Diaries and journals: Collecting insider perspectives in second language research. *The Routledge handbook of research methods in applied linguistics* (s. 348-359). içinde Routledge.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2018). *Designing the user interface: Strategies for effective human-computer interaction.* Pearson.

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

Simmons, N. (2018). Axial coding. M. Allen (Dü.) içinde, *The SAGE Encyclopedia of Communication Research Methods* (s. 80-81). SAGE publications.

Tundjungsari, V. (2020). Mobile learning design using gamification for teaching and learning in algorithms and programming language. *he Challenges of the Digital Transformation in Education: Proceedings of the 21st International Conference on Interactive Collaborative Learning (ICL2018)-Volume 1* (s. 650-661). Springer International Publishing.

Udjaja, Y. (2018). Gamification assisted language learning for japanese language using expert point cloud recognizer. *International Journal of Computer Games Technology*(1), 9085179.

Vlasenko, K. V., Lovianova, I. V., Volkov, S. V., Sitak, I. V., Chumak, O. O., Krasnoshchok, A. V., . . . Semerikov, S. (2022). UI/UX design of educational on-line courses. *CTE Workshop Proceedings*, (s. 184-199).

Vollstedt, M., & Rezat, S. (2019). An introduction to grounded theory with a special focus on axial coding and the coding paradigm. G. Kaiser, & N. Presmeg (Dü) içinde, *Compendium for early career researchers in mathematics education* (Cilt 13, s. 81-100). Springer International Publishing.

Weichbroth, P. (2020). Usability of mobile applications: a systematic literature study. *IEEE Access, 8*, 55563-55577.

Wong, S. K., Nguyen, T. T., Chang, E., & Jayaratna, N. (2003). Usability metrics for e-learning. *On The Move to Meaningful Internet Systems 2003: OTM 2003 Workshops: OTM Confederated International Workshops, HCI-SWWA, IPW, JTRES, WORM, WMS, and WRSM 2003, Catania, Sicily, Italy, November 3-7* (s. 235-252)). Springer Berlin Heidelberg.

Zaharias, P., & Poylymenakou, A. (2009). Developing a usability evaluation method for e-learning applications: Beyond functional usability. *Intl. Journal of Human–Computer Interaction, 25*(1), 75-98.