IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 October 2022, accepted 31 October 2022, date of publication 4 November 2022, date of current version 17 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3219421

RESEARCH ARTICLE

An Empirical Study on the Impact of Gender on Mobile Applications Usability

PAWEŁ WEICHBROTH, (Member, IEEE)
Department of Software Engineering, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 80233 Gdańsk, Poland
e-mail: pawel.weichbroth@pg.edu.pl

ABSTRACT In the area of broadband wireless Internet, mobile applications have already replaced their desktop equivalents and are recognized as valuable tools for any size of businesses and for private use. With the emergence of millions of apps, the quality of their interaction with the user remains an open question for software vendors. While female and male requirements and preferences are not always similar, to the best of our knowledge, few studies have examined the impact of gender on mobile applications usability. Therefore, the goal of this study is to assess their usability from the perspective of female and male users, and to evaluate the differences between them. In our study, based on an experimental setup with a group of 40 users (16 females and 24 males), with regard to three usability attributes, namely efficiency, effectiveness and satisfaction, both qualitative and quantitative data were collected, respectively via pre- and post-testing questionnaires and during application testing sessions, combined with the think aloud protocol. To analyze the collected data, descriptive statistics were extracted from the video data and used to calculate the inferential statistics. With a significance level (alpha) of 5%, our findings show that between the groups of females and males, there were no statistically significant differences in the performance accuracy, average completion time, and perceived satisfaction, since all p values are greater than the assumed alpha. Hence, one can conclude that no effect of gender was observed with regard to the usability of the Gmail application. Overall, the empirical results contribute to the ongoing research on mobile application usability by providing evidence-based insights that we believe may be valuable for both theory and practice.

INDEX TERMS Gender, usability, mobile application, evaluation.

I. INTRODUCTION

With the recent widespread use of smartphones, software vendors' competition to deliver high quality mobile applications is an evident fact today, and by no means a newly discovered one. According to a recent report published by Allied Market Research [1], in 2019 the value of the global mobile application market was estimated at $154.05 billion, up 44.96 percent from 2018 [2]. This positive trend is expected to continue, as the market is projected to reach $407.31 billion by 2026. In particular, consumer spending worldwide in mobile apps reached $83 billion in 2019 [3], and one year later went on to reach a new record of $111 billion [4]. This included spending on subscriptions, premium apps and in-app purchases. However, 95% of Google Play Store apps are still free, along with 90% of iOS App Store apps, ironically generating 98% of most app revenue [5]. Moreover, considering the level of Internet penetration, in the first quarter of 2021, mobile devices (excluding tablets) generated around 55 percent of global website traffic [6].

With over 6.2 billion smartphone users across the world in 2021, the mobile app industry is thriving and projected steadily grow, reaching 7.7 billion in 2027 [7]. It is estimated that an average user in the United States spends 4.23 hours per day with mobile non-voice media [8]. Beyond these, numerous studies show that the voice of users must not be neglected [9], [10], [11], emphasizing the role of usability.

The associate editor coordinating the review of this manuscript and approving it for publication was Eyuphan Bulut.

Besides this, other studies have also shown that mobile application users (hereafter: app users) are particularly inclined to report various types of usability issues, expressing their expectations toward requirements, as well as documenting defects, errors and malfunctions [12], [13], [14]. Eventually, one might conclude that the success (or failure) of an app depends on its perception by the users, whose voices should be listened to and addressed.

The reorientation from desktop to mobile computing has given rise to an entirely new user interface design. Bearing in mind the hardware limitations imposed by mobile devices, including the relatively smaller screen size, it is not just a matter of scaling down a user interface design to develop a mobile-ready application. Therefore, new design patterns have been developed to solve typical issues [15], by redefining and unifying the interaction with the user [16]. However, this unification imposes gender equality, defined here in terms of the state of equal ease of access and use of mobile applications, regardless of gender.

Since, to the best of our knowledge, few studies have recently investigated gender differences in areas related to mobile human-computer interaction, we put forward the following research question: *Does gender impact the usability of mobile applications?* Hence, the goal of this study is to assess the usability from the perspective of female and male users, and to evaluate the differences between them. Along this line of thinking, we aim to start a discussion and self-reflection on that matter, which is not at the forefront of research yet, but is at the core of practice [17], and one of the sustainable development goals [18].

The rest of the paper is organized as follows. Section II discusses the related work devoted to gender issues. Section III describes the research methodology, regarding its theoretical foundations and assumptions, as well as the adopted measurement instruments. Afterward, the experimental design and setup are given in Section IV. The empirical results are presented in Section V which are followed in Section VI by a brief discussion, including the research contributions and limitations. Finally, Section VII summarizes the key findings.

II. RELATED WORK

It seems that there is endless discussion about gender equality and how to achieve it. However, if one takes into account the market of the mobile applications, it seems that these products are gender-neutral. In other words, they have been developed with the intentional assumption of no gender assigned to their users. Nevertheless, the content preferences by gender are very different. For instance, men are likely to use sports, gaming, business news and finance apps [19], while largest women's audience concerns health, shopping, cooking and gardening [20]. Until now, the impact of gender on usability has been the subject of a few studies, regarding different mobile systems and attributes, however with no specific focus on mobile applications.

Střelák et al. [21] examined a mobile augmented reality tourist guide, and found that females were more satisfied with the AR experience compared to males.

A report from a survey, elaborated by Mkpojiogu and Hashim [22], shows that gender had a significant impact on the perceived satisfaction of the usability of mobile banking apps.

Oyibo and Vassileva [23] preformed a study on the four versions of a mobile website. The results show that gender moderated the effect of perceived aesthetics on perceived usability, being stronger for males than for females.

Lim et al. [24], based on empirical studies, found no significant differences when comparing male and female respondents using performance metrics (time on task, errors) from a mobile augmented reality learning environment.

Ibili and Billinghurst [25] evaluated the relationship between the usability of a mobile Augmented Reality (AR) tutorial system and cognitive load. Their findings show that gender did not affect the perceived usefulness, perceived ease of use, and perceived natural interaction. Moreover, gender had no effect on the intrinsic load (cognitive effort), extraneous load (mental effort), or on the germane load (working memory capacity). Interestingly, a strong relationship between the perceived usefulness and the intrinsic load in the group of females, and a strong relationship between the perceived ease of use and the extraneous load in the group of males, were discovered and found significant.

Li and Chen [26] conclude that in the case of learnability, memorability and satisfaction, there was no significant effect of gender in a sample of participants who used a simulated mobile wayfinding application.

To sum up, gender differences, embedded in context of mobile usability, is still at an early stage of research. Nevertheless, at the moment, we know that there is no consensus on this issue, but more arguments are in favour of gender neutrality (see Table 1). Moreover, very little attention has been given to the study of the impact of gender on users' effectiveness and efficiency. In summary, our analysis of recent publications on the topic of gender differences shows that usability research has not yet expanded into mobile technologies. Therefore, this study aims to fill this gap by investigating the effect of gender on mobile applications usability, through the lens of the ISO 9241-11 standard.

III. METHODOLOGY

A. USABILITY DEFINITION

In this study, the definition of usability is borrowed from the ISO 9241-11 standard which states that usability is the ''extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use'' [27]. The rationale behind this choice is the fact that this standard is the most frequently used to define usability in the context of mobile applications [28].

TABLE 1. Summary of related work, regarding studies of the gender effect on mobile systems' usability.

Authors	Year	Sample Size	Research Method	Related Attributes	Gender Effect?
Střelák *et al.* [21]	2016	30 (50%F, 50%M)	Survey	Satisfaction Intuitiveness, Ease of use	Yes* No*
Mkpojiogu and Hashim [22]	2017	150 (n.a.)	Survey	Satisfaction	Yes
Oyibo and Vassileva [23]	2017	n.a.	Survey	Aesthetics	Moderated
Lim *et al.* [24]	2017	46 (50%F, 50%M)	Experiment	Efficiency	No
Ibili and Billinghurst [25]	2019	59 (51%F, 49%M)	Survey	Cognitive load	No
Li and Chen [26]	2020	32 (50%F, 50%M)	Survey	Learnability, Memorability, Satisfaction	No

*: not statistically verified.

TABLE 2. The two observable usability attributes, their definitions and assigned indicators.

Attribute	Definition	Indicators
Observed Effectiveness	the ability of a user to complete a task in a given context	• rate of successful task completion (EFFE1), • total number of steps required to complete a task (EFFE2), • total number of taps related to app usage (EFFE3), • total number of taps unrelated to app usage (EFFE4), • total number of times that the back button was used (EFFE5).
Observed Efficiency	the ability of a user to complete a task with speed and accuracy	• completion time (EFFI1).

TABLE 3. Top three attributes of the perceived usability, their definitions and assigned indicators, or measurement tools, with the predefined rating scales.

Attribute	Definition	Indicators and Measurement Tools	Rating Scale
Perceived Effectiveness	a user's perceived level of workload in a given context	• total number of steps required to complete a task (EFFE2), • total number of taps related to app usage (EFFE3), • total number of taps unrelated to app usage (EFFE4), • total number of times that the back button was used (EFFE5).	7-point Likert scale
Perceived Efficiency	a user's perceived level of application performance	• duration of the application starting (EFFI2), • duration of the application closing (EFFI3), • duration of content loading (EFFI4), • application performance continuity (EFFI5), • duration of the application response to the performed action (EFFI6).	9-point Likert scale
Satisfaction	a user's perceived level of comfort and pleasure	• SUS questionnaire.	5-point Likert scale

B. USABILITY ATTRIBUTES

Keeping the ISO 9241-11 standard in mind, usability is broken down into the following three attributes: effectiveness, efficiency, and satisfaction. Similarly, these three attributes are the top three adapted to evaluate mobile usability [28]. Moreover, their generic (non-specific) nature makes it possible to apply and validate them in any context and in the whole experimental design of this study. Having said that, both the attributes' definitions, along with their corresponding indicators, were adopted from the recent literature concerning contemporary human-computer interaction, embedded in the context of mobile applications [28], [29].

C. ATTRIBUTES CONCEPTUALIZATION

In the light of the results from our latest research [28], two different research methods have been widely used to collect data, namely: controlled observation and survey. However, the quantification can only be performed based on two data sources – the video recordings, or the post-testing questionnaires – for the effectiveness and efficiency attributes. More specifically, the quantitative data is extracted from the video data in order to determine the values of particular indicators (see Table 2) in the case of the former, while in the case of the latter, the user answers a set of questions using a predefined rating scale (see Table 3). Therefore, one should distinguish two different categories of these two attributes: observed, and perceived.

The notions of both observed and perceived effectiveness and efficiency are not very different (compare the definitions given in Table 2 and in Table 3). However, the object of the evaluation is different; while the former category concerns the user's abilities with regard to his/her effectiveness and efficiency related to the performed tasks, the latter reflect the user's perceptions regarding the respectively experienced workload and the application performance. In other words, in the first case, the users' capabilities are evaluated, whereas, in the second case, the users evaluate the interaction quality.

D. ATTRIBUTES OPERATIONALIZATION

As can be seen from reading Table 2, the observed effectiveness and efficiency are directly measured by the five and one quantitative indicators, respectively. On the other

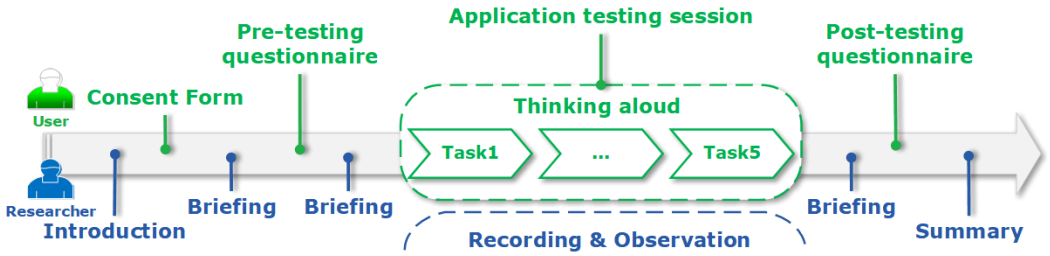

FIGURE 1. Timeline of the data collection procedure with the two roles of the researcher (blue) and the user (green) highlighted, along with the activities performed by each during an application testing session.

hand, Table 3 shows that these two attributes, conceptualized as latent constructs, are manifested by four and five indicators. In particular, both the observed and perceived effectiveness are operationalized by the same four indicators (EFFE2, ..., EFFE5), whereas in the case of efficiency, its observed (EFFI1) and perceived (EFFI2, ..., EFFI6) measurement models do not share any of the indicators.

Satisfaction, the third and final attribute adopted from the ISO 9241–11 standard, is a typical latent variable. In this study, the users' satisfaction was measured by administering Brooke's SUS questionnaire [30] (see Table 5), since it has been widely used for desktop products, and adopted for mobile applications as well [31].

To sum up, it should be noted here that Table 3 is equivalent to the post-testing questionnaire is given in the Appendix B.

E. DATA COLLECTION PROCEDURE

The data collection aims to obtain the primary data that is necessary to:

- describe the profile of the respondents,
- analyze, reproduce and evaluate the interaction, and
- collect the users' feedback, regarding particular usability attributes and their measures.

The following procedure was applied:

1) **Pre-testing questionnaire** is to collect the demographic data, as well as the information regarding the participant's expertise and skills concerning groups of mobile applications along with a usability evaluation with respect to those in daily use (see Appendix A).
2) **Application testing session** is driven by the protocol, including participant recording during application usage, captured by the audio/video hardware apparatus in order to collect both video and voice data. Another method used is participant observation, which refers to a method of generating data that involves observers immersing themselves in a research setting and systematically observing the user's interaction with an application. Moreover, participants are always asked to think aloud about their personal interaction experience (see Subsection IV-B).
3) **Post-testing questionnaire** provides firsthand user feedback regarding the perceived quality of use, conceptualized and operationalized by the usability attributes and their particular measures. Appendix B provides a detailed description of it.

A detailed view of the data collection procedure is depicted on Figure 1.

F. DATA ANALYSIS

By definition, data analysis is fundamentally an iterative process in which a researcher extracts the premises necessary to formulate conclusions [32]. In this context, the data analysis process involves: (a) video content analysis (VCA), b) documenting all identified application errors, defects and malfunctions, and c) extracting all numerical values required to estimate usability attribute indicators. It should be noted here that it is a common practice to use a video player application, or any other software tools which support this process. Apart from this, a variety of visualization techniques are usually utilized to facilitate analysis and interpretation of the obtained results. For instance, a timeline is a graphical method of displaying a list of a user's actions in chronological order.

In the last stage, eventually, the results are summarized and interpreted in terms of usability evaluation. In the short version, the report is typically divided into four sections: (1) bugs and errors, (2) design, (3) performance, and (4) findings and recommendations. However, if needed, a full version might also include: (5) respondents' profiles, (6) research goals and methodology, (7) disclaimers, and (8) other additional information, as long as they are relevant to its receivers.

However, due to the limitations of this paper, the further reporting is actually based only on information that is strictly related to the research question. The information reporting was therefore intentionally limited and deliberately focused on the specific usability facets.

IV. EXPERIMENTAL SETUP

A. RESEARCH APPARATUS

For real-time image and voice capture and recording, we used a Genee Vision 150 document camera. It was connected by a USB port to a laptop, required minimal set-up, and saved audio-video data on a local hard drive in the *avi* file format. During a test, the device was positioned on the desk next to the user and directed toward the smartphone screen. Its physical

location did not result in any obstacles or obstruction in the user's sitting position or testing performance. For each test, an explicit confirmation of the user was requested.

B. TESTING SESSION CONFIGURATION

Despite the wealth of alternatives, the unmoderated user session approach [33] was adopted and adapted in this study, however with two distinctions. First, the facilitator was present during a testing session, and second, the sessions were conducted in the same laboratory (both the user's and a facilitator's physical location were the same). It is worth noting here that an unmoderated user session is recommended when the focus of the study concerns a few specific elements, rather than a general review [34].

The usability testing session was performed in two-rounds. In the first round, a user could freely select a mobile application for usability testing. However, the user must confirm having using this app for at least three months to proceed. This assumption simply aimed to eliminate the long tail of questions and answers from the observer to the user, and vice-versa. In other words, the first round was assumed to be a warm-up, involving gentle exercise and relaxation.

In the second round, the participant was asked to perform a series of tasks, prescribed in points and given to the user on a piece of paper. Before the participants started the first round, they were informed of the purpose of the study and given a brief description of it. Firstly, a link to an electronic pre-questionnaire was administered to the participant. Next, the actual testing session was performed, and recorded using an external camera device. Afterward, a link to the electronic post-questionnaire was submitted to the user. The short summary served as the closing theme.

C. PARTICIPANTS

The purposive (non-probability) sampling technique was used to determine the participants from the population due to its cost efficiency. In other words, it was the deliberate choice of the mobile application users who volunteered to participate in the study in response to a request sent by email.

The participants were recruited among computer science students, with an average age of 23.58 years ($\pm$3.11), sex (60 percent males, and 40 percent females), professional background (ranging from 0 to 10 years), and varied in their use of different mobile devices (smartphones, tablets and ebook readers), with different mobile operating systems (Android, and Apple iOS).

D. MOBILE DEVICES

Considering the market of mobile devices, smartphones are the most widely used handheld computers. Therefore, it was the obvious choice as the testing platform, since familiarity and routine in daily use is relatively high among the youth. Moreover, having in stock a new smartphone, we always asked the user to use their own smartphone during the testing to preserve the convenience and comfort.

All participants were using their own devices, having a minimum screen size of 5 inches. It is worth noting here that before a session, the mobile application being the subject of the study was optionally updated, and each device was required to be restarted to remove all running applications in the background, to preserve its efficiency.

E. TIME FRAMES

Between January and May 2019, a series of usability testing sessions on the Gmail mobile application were performed. In total, the data from 44 sessions were collected, including 40 valid and 4 invalid sessions, due to the incompleteness of the content, or the lack of understanding demonstrated by individual participants during the testing sessions.

F. USABILITY TESTING PROCEDURE

An assessment research design was adopted in the extent of the usability testing method [35]. In particular, a fixed and rigid testing procedure was developed, which consisted of five independent tasks:

1) Send an email message.
2) Forward an email message.
3) Delete an email message.
4) Archive an email message.
5) Mark an email message with a star.

There was no time limit and no limit for the maximum number of gestures to undertake by a user, who was also allowed to ask questions, and encouraged to thinking aloud while carrying out the given tasks. In the case of the first and second task, the name of the email recipient was given, while the former task also required a title consisting of 4 letters, as well as a short message of 4 letters (one word).

G. COLLECTED DATA

On average, a single session lasted approximately 15 minutes, while the duration of the application testing session was up to 5 minutes. The volume of collected data covers 40 recordings with an estimated size of approx. 10 gigabytes. The collected data covers 40 participants (16 females and 24 males), and they were used for analysis without any missing values.

Google Forms was used to collect data from the users regarding both the pre- and post-questionnaire items, due to its simple setup and user-friendly interface. Table 12 presents the collected data regarding the perceived effectiveness and efficiency, while Table 13 shows the raw data collected via the SUS questionnaire, and the calculated SUS scores, along with assigned scales and their verbal interpretations.

H. DATA ANALYSIS

For the purpose of data analysis, we first inspected the video content that comprises annotation procedures in which the user's actions and application responses are identified and placed on a timeline. Secondly, we documented all identified application defects and errors, along with all issues verbally identified during the sessions and reported in the post-testing

questionnaires. Thirdly, we extracted the numerical values necessary to calculate usability attribute indicators in the following way.

The task performance reconstruction was performed based on the video data analysis, supported by the RVDA tool (Retrospective Video Data Analyzer). This tool was designed to extract and annotate time series events. A detailed description of the RVDA application can be found in [29].

In the first run, if a user accomplished a task successfully, the EFFE1 indicator was assigned a value of 1, and if the user failed, a value of 0 was assigned. In the second run, the values of the remaining four indicators were individually extracted if the user successfully performed the test.

Table 4 presents the result of the extracted data from a single video recording for Task2, performed by the user coded as R1. Considering the user's effectiveness, the reconstructed interaction shows that performing the task of forwarding an email message required seven steps (first column), involving a total of 26 taps on the screen (the sum of the *effe*3 values of the third column gives the value of the EFFE3 indicator).

TABLE 4. Reconstructed Task2 performance of the R1 user.

Step	Action	effe3	effe4	effe5
1	Execute application	1	0	0
2	Move screen	2	0	0
3	Choose an email message	1	0	0
4	Select Hamburger Menu	1	0	0
5	Select Forward Option	1	0	0
6	Type an email address	19	0	0
7	Send	1	0	0

Afterward, the values of the four observed effectiveness indicators were individually calculated, based on the values from the first, third, fourth, and fifth column, respectively. The results for all users who accomplished Task2 successfully are given in Table 7. The remaining data, related to the other tasks, are given in Tables 6 to 10.

It is worth noting here that any user tap on the screen that unintentionally deviated from what was required has been classified as unrelated to app usage, and marked with the *effe*4 pin; for instance, a tap on the wrong user interface element.

Regarding the observed efficiency, time to completion (EFFI1) is the calculated amount of time required by a user for any particular task to be completed. For each task, all EFFI1 values were individually extracted by estimating the difference between the end time and the start time. The summary for all five tasks is given in Table 11.

Finally, the quantitative data was analyzed using mean scores, percentage, standard deviation values, as well as other statistics, and eventually triangulated with the qualitative data for interpretation and conclusion.

To collect and organize the data, we used spreadsheet software, and for statistical analysis, we used *jamovi* [36], which is built on top of R, with a library of additional modules, and under active development by the scientific community. Descriptive statistics were extracted from the video data and used to calculate the inferential statistics. For this purpose, Student's *t*-test was employed to compare the means of particular indicators for the two independent groups of females and males, or the Mann-Whitney U-Test, for the indicators that did not follow a normal distribution. The level of significance (α) was set equal to 5%.

V. RESULTS

A. EFFECTIVENESS EVALUATION

The video content analysis showed that all tasks were accomplished by both groups of participants. Therefore, the rate of successful task completion (EFFE1) was 100% for both the females and males. Being the primary measure of the observed effectiveness, this shows the degree of completeness in achieving certain goals by the users.

However, in investigating the task completion times (see Table 11), we shall further evaluate only the first task, because the remaining tasks, being relatively short, were performed very similarly with regard to the results of the observed effectiveness indicators applied.

Table 14 shows the statistics of the remaining four observed effectiveness indicators, estimated for Task1. It is interesting to note the strong similarity in the performance while comparing the females with the males when looking at the average (*M*), and later at the mean difference (*MD*). In other words, the number of steps required to complete the task (EFFE2), along with the number of taps related to app usage (EFFE3) on the one hand, as well as the number of taps unrelated to app usage (EFFE4) and the back button usage (EFFE5), were not very different. The box-and-whisker diagrams depicted in Figure 2 show the locality, spread and skewness of these four indicators through their quartiles, indicating variability outside the upper and lower quartiles; moreover, the estimated average is included and marked by the *x* symbol.

The total number of errors was lower in the case of the females group than in the males group. All of the errors were similar and were caused by tapping an incorrect area of the screen, consequently requiring pressing the back button to return to the previous screen. Therefore, these errors were not severe and the users easily recovered from them. In other words, the occurrence of these errors did not denote a failure in understanding the administered instructions by the participants. In any case, test if the differences are significant, it is necessary to apply a relevant statistical test.

Since Student's t-test is parametric and assumes normality of the data and equality of variances across comparison groups, the Shapiro-Wilk test is used to check if the effectiveness indicators follow a normal distribution. The null hypothesis (H0) states that the indicator is normally distributed, and the alternative hypothesis (H1) states the opposite.

The Shapiro-Wilk test was performed, and showed (see Table 14) that the distributions of the three indicators (EFFE2, EFFE4, and EFFE5) for both groups departed significantly from normality (*p*-value is lower than 0.001 for all three). Based on this outcome, a non-parametric test was used, namely the Mann-Whitney U-Test. In this case, under H0, the

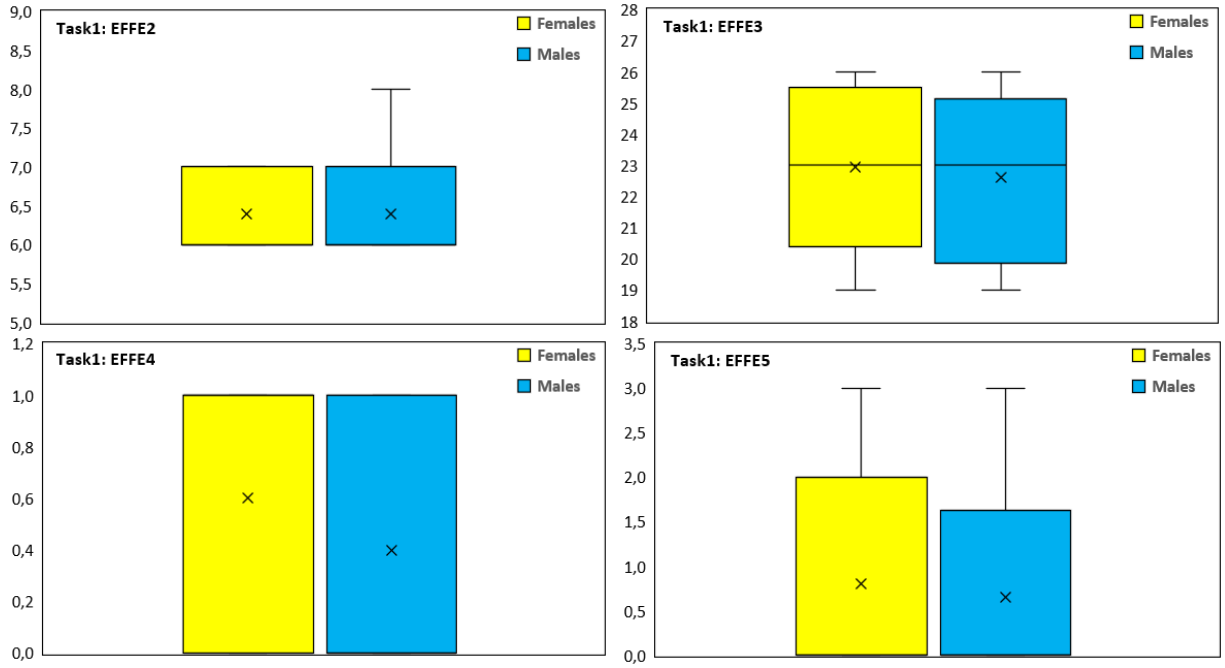

FIGURE 2. Box-plots of the four observed effectiveness indicators for Task1 for the groups of females (yellow) and males (blue).

distributions of both groups are identical, whereas H1 claims the opposite.

The Mann-Whitney U-Test (MWU) showed that these differences were not statistically significant for each of the three indicators (EFFE2 $U = 175$ and $p = 0.533$, EFFE4 $U = 152$ and $p = 0.194$, and for EFFE5 $U = 168$ and $p = 0.421$).

For the EFFE3 indicator, while the data distributions for both groups can be assumed to be normal ($p = 0.190$ for the females group, and $p = 0.058$ for the males group), and have the same variance (the F-test of equality of variances gives $F = 0.208$ and $p = 0.651$), we performed an Independent Sample T-test to compare the means. The estimated Student's test p-value is greater than 0.05 ($p = 0.657$), and Cohen's d is classified as a very small effect size, which confirm that there were no significant differences between these two groups.

To sum up. Since we failed twice to reject the null hypothesis, using both non-parametric and parametric testing, there is insufficient evidence to say that the observed effectiveness in performing Task1 in the females group was different from the males group. In other words, this evidence supports the hypothesis that **no effect of gender** was observed in this sample.

Table 15 presents the results of the questionnaire carried out after application testing, regarding all five tasks. The responses from the post-test questionnaire indicate that the participants were generally in favor of the designed interface, as well as the features, that all together facilitate efficient task performance. In this regard, the number of taps related to app usage was found to be highly satisfactory. On the other hand, users highly rated the lack of taps unrelated to app usage. Hence, we might conclude that the size of the touch controls were scaled up appropriately to the size of the user interface.

The distributions of the collected data with respect to perceived effectiveness indicators are not normal, since the p values are less than 0.05. Thus, to examine whether there was an effect of gender, one should perform the MWU test. The four tests involved the calculation of the U statistic and p-value for each indicator, independently and respectively for each group.

The outcomes show that there is insufficient evidence to reject the null hypothesis for the three indicators which means that the perceived effectiveness, described by the number of taps related (unrelated) to app usage, as well as by the number of times the back button was used, was not found to be statistically significantly different between the females and males (EFFE3 $U = 166$ and $p = 0.448$, EFFE4 $U = 165$ and $p = 0.411$, and EFFE5 $U = 168$ and $p = 0.452$). However, in the case of the EFFE2 indicator, denoting the number of steps required to complete a task, the estimated values of $U = 125$ and $p = 0.04$ show that the difference is statistically significant between females ($Mdn = 6$) and males ($Mdn = 5$).

B. EFFICIENCY EVALUATION

We start by providing an analysis and interpretation of the results in Table 16. The averages of the estimated task completion time (EFFI1) for all five tasks were roughly the same for the females and males (see Figure 3). In the observable sample, one can notice that on average the females required less time to accomplish Task1, Task2 and Task5. But, to find out whether the difference is significant, it is necessary to apply a relevant statistical test.

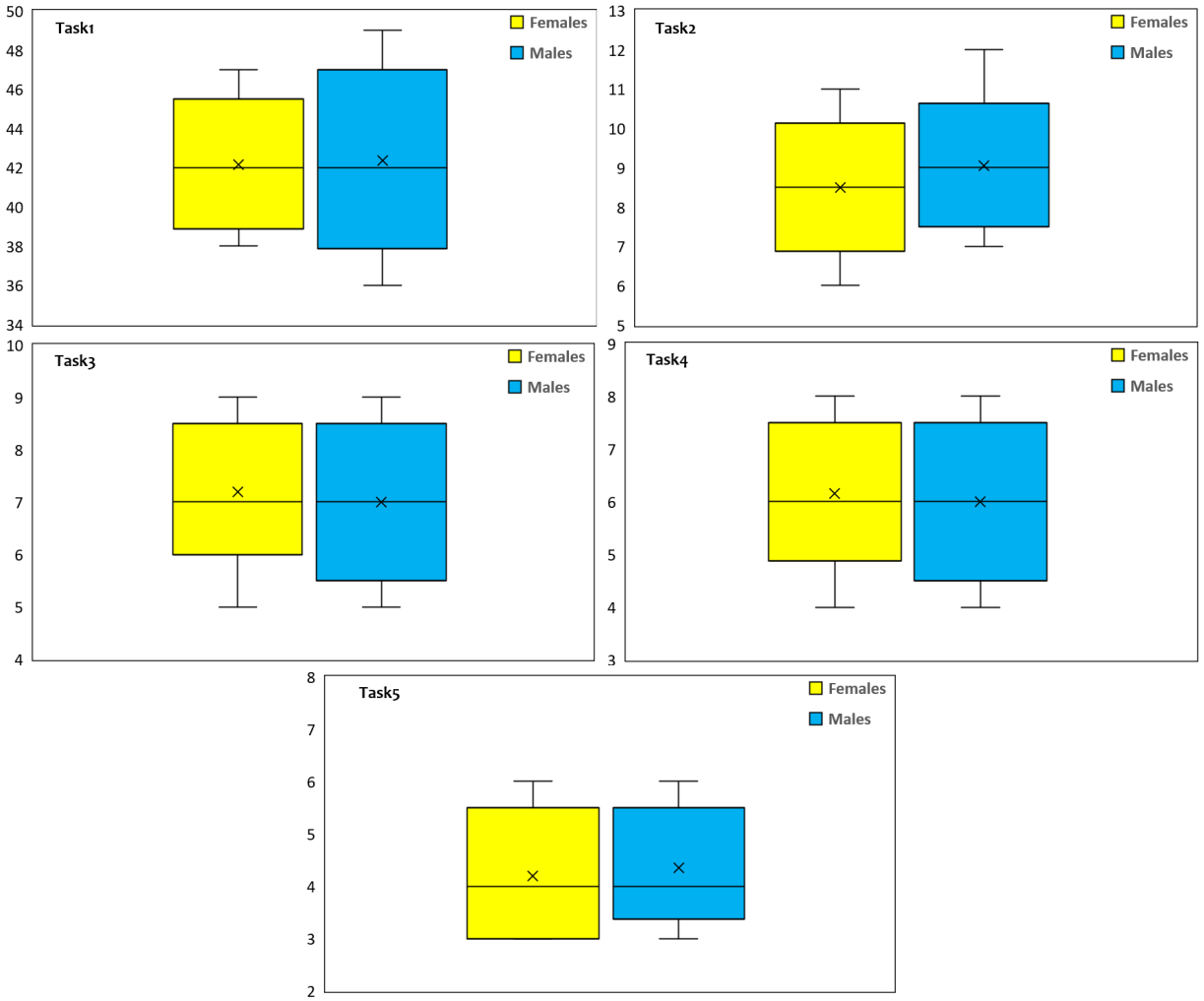

FIGURE 3. Box plots of the times to completion for all five tasks for both females (yellow) and males (blue).

The distributions of completion times for both groups can be assumed in all tasks to be normal (the Shapiro–Wilk test gives an observed probability level of the p-value greater than 0.05), and to have the same variance (the F-test of equality of variances also gives a p-value greater than 0.05), with the exception of Task5 (the Shapiro–Wilk test gives a p-value of 0.026 for the females and a p-value of 0.07 for the males, however the F-test of equality of variances gives a p-value of 0.453). Therefore, we performed an Independent Sample T-test to compare the means.

The mean comparison between the results of the females and males shows that the two means are not significantly different in each task (see Table 16). This is because for p, the value is greater than 0.05 (significance level). Therefore, in each test, we cannot reject the null hypothesis, and we have to conclude that **between the groups of females and males, there were no significant differences in the average completion time**. Moreover, the estimated Cohen's ds also indicated small or very small differences, using the classification further elaborated by Cohen [37] and expanded by Sawilowsky [38].

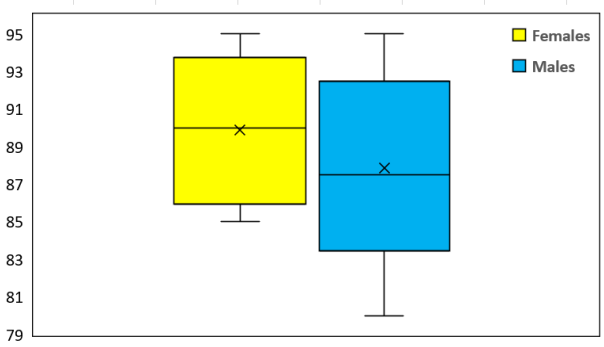

FIGURE 4. Box plot of the perceived satisfaction, reflected by the estimated SUS scores, for the group of females (yellow) and the group of males (blue). The average is depicted by the x symbol for each group.

On the other hand, all respondents highly evaluated the perceived efficiency of the testing application (see Table 17), measured by the five indicators that gained average scores of 7.5 or higher (using the 9-point Likert scale). Taken at face value, one can conclude that the application performance

TABLE 5. System usability scale [30].

Code	Item	1	2	3	4	5
Q1	I think that I would like to use this system frequently					
Q2	I found the system unnecessarily complex					
Q3	I thought the system was easy to use					
Q4	I think that I would need the support of a technical person to be able to use this system					
Q5	I found that the various functions in this system were well integrated					
Q6	I thought there was too much inconsistency in this system					
Q7	I imagine that most people would learn to use this system very quickly					
Q8	I found the system very cumbersome to use					
Q9	I felt very confident using the system					
Q10	I needed to learn a lot of things before I could get going with this system					

met their requirements with regard to the data processing duration. It is worth noting here that the performance of mobile email client applications strongly relies on the Internet connection speed, therefore imposing a maximum possible limit on the user's abilities.

Considering whether there were (or not) differences in the observed efficiency between the groups of females and males, the Shapiro-Wilk tests show that the data distributions are not normal since all p-values are lower than 0.05. Due to this fact, the MWU tests were performed and investigated. The results are consistent such that in the case of the all indicators, there is insufficient evidence to reject the null hypothesis, since all p-values are greater than 0.05, which means that the differences are not statistically significant. Similarly, **no effect of gender** has been recognized with respect to the perceived efficiency.

C. SATISFACTION EVALUATION

The average SUS score for the females group was 89.4 ($SD =$ 3.48), while for the males group, it was 87.7 ($SD = 3.68$). It is worth noting that all calculated SUS scores were above 70 points which means that for all users the level of perceived usability was acceptable when using the adjective rating scale elaborated by Bangor et al. [39] (see Figure 5).

With regard to the grade scale and adjective rating, 9 females and 10 males (in total 47.5%) scored the app as grade **A** (the top of the scale), while the rest (7 females and 14 males) as grade **B**. With SUS scores between 85 and 99, 87.5% of respondents (all females and 19 of 24 males) rated the app as *excellent*, while the rest considered the app to be *good*, with a minimum score of 80 points.

One can notice that the SUS score was higher for the females group, indicating a mean difference of 1.67 points, compared with the males group (see Figure 4). Nevertheless, to determine whether the difference is significant, it is necessary to apply a relevant statistical test. The distributions of scores for both groups (females, and males) can be assumed to be normal since the Shapiro–Wilk test gives p-values of 0.100 and 0.086, respectively, and to have the same variance, where the F-test of equality of variances gives a p-value of 0.834. All p-values are greater than 0.05, therefore, an Independent Sample T-test is valid to compare the means.

The mean comparison (see Table 18) of perceived satisfaction declared by the females and males shows that the means of the SUS scores are not significantly different because the p value is greater than 0.05 ($t(38) = 1.43$, and $p = .160$). Finally, one can conclude that we cannot reject the null hypothesis. This finding led to the conclusion that the SUS scores did not significantly differ based on the respondents' sex, and we can assume that **between the groups of females and males, there were no significant differences in the perceived satisfaction** of the tested object. Moreover, the estimated value of Cohen's d of 0.463 also indicates small differences between the estimated means.

An analysis of the extreme low and high SUS scores showed that none of the users considered the service to be "*best imaginable*," with a perfect 100 SUS score. Only one respondent (male) gave a SUS score of 80 points, while, on the other hand, three ratings (2 females, and 1 male) were 95 points each. These results and the overall average rating show that there is little room for improvement, and no need to address critical usability issues of the Gmail mobile application.

VI. DISCUSSION

One can notice that to collect data, both explicit (questionnaire, thinking aloud) and implicit (audio-video recording) methods were used. On the other hand, both types of data, namely quantitative and qualitative, were collected, extending the evaluation capabilities. We argue that by performing and combining these methods, a unified view of the quality of use can be achieved. In other words, usability evaluation is undertaken by using both objective (evidence-based measures) and subjective (users' judgment) information sources.

The thinking aloud protocol was applied since we wanted to know what a user thinks about the application design, performance, and overall interaction, and we wanted to identify his/her feelings, perceptions, and emotions. Interestingly, our respondents were very focused on the task's performance, obstructing their verbal communication capacity. Since a single testing session did not last for more than 5 minutes on the one hand, and the respondents were familiar with the Gmail application on the other, we obtained a very small amount of qualitative data from the thinking aloud protocol. In conclusion, in our opinion, this method could bring more valuable information in the case of testing newly developed user interface designs, and during unmoderated testing sessions.

A. THEORETICAL CONTRIBUTIONS

Our study makes several theoretical contributions to the literature. First, we argue that there are not significant differences in the average completion times in performing the tested tasks by females and males, as well as in the performance accuracy and perceived usability. Reflecting on our research question, no effect of gender was observed. However, other authors argue that gender differences affect user preferences [40], or report the presence of differences in the usability of the interface design of car navigation systems [41].

Other related studies also claim that gender should be taken into account in designing mobile application interfaces [42]. Therefore, it is worth noting here that the relationship between gender and usability seems not to have reached a collective theoretical consensus regarding different areas of human-computer interaction.

B. PRACTICAL CONTRIBUTIONS

From the practical perspective, gender is considered a determinant of user behavior across domains such as on-board navigation systems [41], online auctions [43], on-line shopping [44], and Internet blogs [45] as well as marketing [46]. It is also a sociocultural factor that impacts the planning of business strategies [47], and well-being in the work environment [48], where the male–female division is taken to be the most basic human characteristic [40].

Now, taking into account the context of our study, and considering the results from the aforementioned studies, one might ask: do we need to develop two versions of an email client mobile application – one for females, and one for males? Since there are no significant differences reported according to gender, the answer is negative. However, it does not mean that further studies should not be undertaken to confirm (or not) the obtained results.

C. THREATS TO VALIDITY

There are several limitations of this research that should be noted. Firstly, the sample population (i.e. computer science students) only represents a relatively small subgroup of end-users, considering both their number and age, as well as culture, language, and lifestyle. Therefore, the findings from the usability test may not be fully representative of the general user community. Further research should investigate different groups of users according to their age, profession, and attitudes. Performing these tests in different contexts and settings will enable the findings to be generalized by gathering more evidence-based conclusions.

Secondly, the sample is unbalanced between the two groups (the sample set contains 40% females vs 60% males), which makes the comparison of these two groups more difficult. However, its effect may be treated as small, if one takes into account the difference of 8 users. Thus, gender-balanced groups are needed to obtain more representative results. Moreover, the use of multiple bivariate tests may have contributed to an inflated chance of spurious significant findings. In other words, these tests offer no control of confounding variables and increase the risk of inflating Type I errors.

Thirdly, the study suffers from testing and evaluating only one mobile application. The rationale behind its choice requires more elucidation. Thus, more follow-up data are still required, including not only matured products, but also high-fidelity prototypes. Hence, future research should be addressed to other applications, possibly having other functionalities, or being at different stages of development. We hope that our research approach will find followers, but also that these followers will perform further studies by incorporating other methods (e.g. cognitive walkthrough, eye tracking) so far not considered in our approach.

In this line of thinking, fourthly, the context-dependent indicators, reflecting specific application features related to tasks given to a user to perform, were not defined, but could have revealed differences in both observed and perceived usability. In other words, the level of understanding the tasks, and, as a consequence, the capabilities to perform them, might have affected the user's performance, as well as his/her perceptions of the interaction quality.

Moreover, fifthly, analogous indicators were used to measure and evaluate both observable and perceived usability. While their theoretical foundations are strongly rooted in the state-of-the art literature, the reliability of two measurement tools used to evaluate the perceived effectiveness and efficiency were not assessed. Thus, future research should also cover this facet, since it is important to fully assess how ''good'' these scales are at measuring these two attributes.

Sixthly, qualitative methods often bring into question the credibility of the results obtained [49]. Indeed, the data analysis concerned a relatively small sample size, and was performed by one investigator, however facilitated by the RVDA tool. This software was designed and developed for video content analysis, including such operations as: extracting, annotating and reasoning about time and events. On the other hand, the existing functionality exhibits good capabilities, while the required workload did not exceed one hour in the case of video data up to 5 minutes long.

VII. CONCLUSION

Through the lens of existing research instruments and methods, adopted from modern usability theory and practice, the current study advocates gender equity with respect to mobile applications usability. Drawing upon the ISO 9241-11 standard, and more specifically, considering three usability attributes, namely effectiveness, efficiency and satisfaction, no significant differences were found between females and males.

These results contribute to the on-going human-computer interaction research, in particular regarding the mobile applications domain, which has recently attracted and gained considerable attention from both academia and the software industry. Nevertheless, more research is needed to provide more evidence in this matter. Since the global market of

TABLE 6. The calculated values of the observed effectiveness indicators for Task1.

ID	Sex	EFFE1	EFFE2	EFFE3	EFFE4	EFFE5
R1	M	1	6	19	0	0
R2	M	1	6	21	0	0
R3	M	1	6	20	1	0
R4	F	1	6	19	0	0
R5	F	1	6	21	1	0
R6	M	1	6	20	1	0
R7	F	1	7	23	0	1
R8	M	1	6	26	0	3
R9	F	1	7	22	1	0
R10	M	1	6	23	0	0
R11	F	1	6	23	0	0
R12	M	1	6	24	1	0
R13	M	1	6	25	0	0
R14	F	1	6	22	1	0
R15	M	1	6	21	1	0
R16	M	1	6	19	0	0
R17	M	1	6	22	1	0
R18	M	1	6	25	0	1
R19	F	1	6	23	1	0
R20	M	1	6	19	0	0
R21	M	1	6	24	1	1
R22	F	1	6	23	0	0
R23	M	1	6	22	1	0
R24	M	1	6	26	0	2
R25	M	1	6	24	0	0
R26	M	1	6	23	0	0
R27	F	1	7	26	1	3
R28	F	1	6	25	1	1
R29	M	1	7	23	0	0
R30	F	1	6	22	1	1
R31	M	1	7	19	0	0
R32	F	1	6	19	0	0
R33	M	1	6	25	0	1
R34	F	1	7	26	0	2
R35	M	1	8	23	0	0
R36	F	1	7	21	1	0
R37	F	1	6	26	0	0
R38	M	1	7	25	0	1
R39	F	1	6	25	0	1
R40	M	1	7	23	0	0

TABLE 7. The calculated values of the observed effectiveness indicators for Task2.

ID	Sex	EFFE1	EFFE2	EFFE3	EFFE4	EFFE5
R1	M	1	7	26	0	0
R2	M	1	7	26	0	0
R3	M	1	7	27	0	0
R4	F	1	7	26	0	0
R5	F	1	8	30	0	0
R6	M	1	7	28	0	0
R7	F	1	7	29	0	0
R8	M	1	7	30	0	0
R9	F	1	7	28	0	0
R10	M	1	7	27	0	0
R11	F	1	8	32	2	0
R12	M	1	8	31	1	0
R13	M	1	7	26	0	0
R14	F	1	7	28	0	0
R15	M	1	7	28	0	0
R16	M	1	7	28	0	0
R17	M	1	7	29	0	0
R18	M	1	7	30	1	0
R19	F	1	8	30	1	0
R20	M	1	8	32	2	0
R21	M	1	7	30	0	0
R22	F	1	7	29	0	0
R23	M	1	7	28	0	0
R24	M	1	7	29	0	0
R25	M	1	7	26	0	0
R26	M	1	7	27	0	0
R27	F	1	7	28	0	0
R28	F	1	7	29	0	0
R29	M	1	7	30	0	0
R30	F	1	7	30	0	0
R31	M	1	7	29	0	0
R32	F	1	7	28	0	0
R33	M	1	7	27	0	0
R34	F	1	7	28	0	0
R35	M	1	7	28	0	0
R36	F	1	8	32	0	0
R37	F	1	7	28	0	0
R38	M	1	7	29	0	0
R39	F	1	7	26	0	0
R40	M	1	7	27	0	0

mobile devices is still growing, one might expect that the studies of the gender effect will find followers.

When evaluating the usability of mobile applications, one can consider a spectrum of different methods to be applied, including those both quantitative and qualitative in nature. Their selection and adoption is usually determined by the attributes of the users of the system (e.g. age, profession, disabilities) on the one hand, and based on the cost, duration, and available resources on the other.

On the other hand, future studies can also cross the border of the usability realm. Other possible areas to investigate concern the user experience and affective computing. While the former includes the user's beliefs, perceptions, and preferences, the latter recognizes, interprets, processes, and simulates the user's affects. Undeniably, such research directions can bring a better understanding of human computer interaction, contributing by bringing new insights for user interface designers.

APPENDIX A
PRE-TESTING QUESTIONNAIRE
A. INTRODUCTION

The pre-testing questionnaire was used to collect demographic data, and information regarding the participant's expertise and skills concerning groups of mobile applications along with a usability evaluation with respect to those being in daily use.

Q1. What is your age?

Q2. What is your gender?

☐ Female.
☐ Male.

Q3. What is the level of your education?

Q4. What is your professional experience (in years)?

Q5. Please indicate the type of mobile device and the operating system (OS) you have been using for at least 3 months:

TABLE 8. The calculated values the observed effectiveness indicators for Task3.

ID	Sex	EFFE1	EFFE2	EFFE3	EFFE4	EFFE5
R1	M	1	3	3	0	0
R2	M	1	3	3	0	0
R3	M	1	3	3	0	0
R4	F	1	3	3	0	0
R5	F	1	3	5	2	0
R6	M	1	3	5	2	0
R7	F	1	3	3	0	0
R8	M	1	3	3	0	0
R9	F	1	3	3	0	0
R10	M	1	3	4	1	0
R11	F	1	3	5	2	0
R12	M	1	3	4	1	0
R13	M	1	3	3	0	0
R14	F	1	3	3	0	0
R15	M	1	3	3	0	0
R16	M	1	3	3	0	0
R17	M	1	3	4	1	0
R18	M	1	3	3	0	0
R19	F	1	3	3	0	0
R20	M	1	3	3	0	0
R21	M	1	3	3	0	0
R22	F	1	3	4	1	0
R23	M	1	3	5	2	0
R24	M	1	3	4	1	0
R25	M	1	3	4	1	0
R26	M	1	3	3	0	0
R27	F	1	3	3	0	0
R28	F	1	3	3	0	0
R29	M	1	3	3	0	0
R30	F	1	3	3	0	0
R31	M	1	3	3	0	0
R32	F	1	3	3	0	0
R33	M	1	3	4	1	0
R34	F	1	3	5	2	0
R35	M	1	3	4	1	0
R36	F	1	3	3	0	0
R37	F	1	3	3	0	0
R38	M	1	3	3	0	0
R39	F	1	3	3	0	0
R40	M	1	3	3	0	0

TABLE 9. The calculated values of the observed effectiveness indicators for Task4.

ID	Sex	EFFE1	EFFE2	EFFE3	EFFE4	EFFE5
R1	M	1	3	3	0	0
R2	M	1	3	3	0	0
R3	M	1	3	3	0	0
R4	F	1	3	3	0	0
R5	F	1	3	4	1	0
R6	M	1	3	3	0	0
R7	F	1	3	3	0	0
R8	M	1	3	3	0	0
R9	F	1	3	4	1	0
R10	M	1	3	4	1	0
R11	F	1	3	3	0	0
R12	M	1	3	3	0	0
R13	M	1	3	3	0	0
R14	F	1	3	3	0	0
R15	M	1	3	3	0	0
R16	M	1	3	5	2	0
R17	M	1	3	5	2	0
R18	M	1	3	3	0	0
R19	F	1	3	4	1	0
R20	M	1	3	3	0	0
R21	M	1	3	4	1	0
R22	F	1	3	3	0	0
R23	M	1	3	4	1	0
R24	M	1	3	3	0	0
R25	M	1	3	3	0	0
R26	M	1	3	3	0	0
R27	F	1	3	3	0	0
R28	F	1	3	3	0	0
R29	M	1	3	4	1	0
R30	F	1	3	4	1	0
R31	M	1	3	4	1	0
R32	F	1	3	3	0	0
R33	M	1	3	3	0	0
R34	F	1	3	3	0	0
R35	M	1	3	3	0	0
R36	F	1	3	3	0	0
R37	F	1	3	3	0	0
R38	M	1	3	3	0	0
R39	F	1	3	4	1	0
R40	M	1	3	5	2	0

Device / OS	Android	iOS	Windows OS	Other
Tablet				
E-book reader				
Smartphone				

Q6. What mobile apps have you been using for more than three months:

1) Web Browsers:
   - ☐ Chrome
   - ☐ Firefox
   - ☐ Safari
   - ☐ Other: specify: .....
2) Email Clients:
   - ☐ Gmail
   - ☐ Outlook
   - ☐ Spark
   - ☐ Other: specify: .....
3) Social Media:
   - ☐ Facebook
   - ☐ Instagram
   - ☐ LinkedIn
   - ☐ Other: specify: .....
4) Instant messaging:
   - ☐ Facebook Messenger
   - ☐ Snapchat
   - ☐ WhatsApp
   - ☐ Other: specify: .....
5) Entertainment:
   - ☐ YouTube
   - ☐ Netflix
   - ☐ Spotify
   - ☐ Other: specify: .....

Q7. How often do you use mobile applications?

- ☐ every day – more often than every hour (>16),
- ☐ every day – once an hour on average (<16),
- ☐ every day – occasionally (<4),
- ☐ once a day,

TABLE 10. The calculated values of the observed effectiveness indicators for Task5.

ID	Sex	EFFE1	EFFE2	EFFE3	EFFE4	EFFE5
R1	M	1	2	2	0	0
R2	M	1	2	2	0	0
R3	M	1	2	2	0	0
R4	F	1	2	2	0	0
R5	F	1	2	2	0	0
R6	M	1	2	2	0	0
R7	F	1	2	2	0	0
R8	M	1	2	2	0	0
R9	F	1	2	2	0	0
R10	M	1	2	3	1	0
R11	F	1	2	3	1	0
R12	M	1	2	3	1	0
R13	M	1	2	3	1	0
R14	F	1	2	2	0	0
R15	M	1	2	2	0	0
R16	M	1	2	2	0	0
R17	M	1	2	2	0	0
R18	M	1	2	2	0	0
R19	F	1	2	2	0	0
R20	M	1	2	3	1	0
R21	M	1	2	2	0	0
R22	F	1	2	2	0	0
R23	M	1	2	2	0	0
R24	M	1	2	2	0	0
R25	M	1	2	2	0	0
R26	M	1	2	2	0	0
R27	F	1	2	2	0	0
R28	F	1	2	2	0	0
R29	M	1	2	3	1	0
R30	F	1	2	3	1	0
R31	M	1	2	3	1	0
R32	F	1	2	2	0	0
R33	M	1	2	2	0	0
R34	F	1	2	2	0	0
R35	M	1	2	2	0	0
R36	F	1	2	2	0	0
R37	F	1	2	2	0	0
R38	M	1	2	2	0	0
R39	F	1	2	2	0	0
R40	M	1	2	2	0	0

TABLE 11. The time completion (EFFI1) calculated for all five tasks.

ID	Sex	Task1	Task2	Task3	Task4	Task5
R1	M	37	9	6	8	4
R2	M	36	10	5	6	3
R3	M	40	11	7	5	4
R4	F	39	7	8	5	3
R5	F	41	8	7	4	4
R6	M	42	12	8	7	5
R7	F	43	6	6	5	5
R8	M	38	9	9	5	6
R9	F	41	9	8	6	6
R10	M	40	11	8	6	6
R11	F	39	10	7	7	4
R12	M	46	10	7	6	5
R13	M	45	9	6	7	4
R14	F	44	8	6	6	5
R15	M	39	8	7	6	3
R16	M	43	7	8	7	3
R17	M	44	8	7	7	4
R18	M	48	9	6	6	5
R19	F	47	7	7	8	3
R20	M	40	10	6	7	5
R21	M	39	8	7	8	6
R22	F	38	9	7	6	4
R23	M	45	8	6	5	5
R24	M	42	9	5	6	4
R25	M	49	9	6	7	3
R26	M	43	8	7	8	3
R27	F	44	10	8	7	5
R28	F	42	11	9	8	3
R29	M	46	7	8	6	4
R30	F	45	8	8	7	4
R31	M	42	9	9	5	5
R32	F	43	9	7	5	5
R33	M	39	7	7	5	6
R34	F	40	10	8	6	3
R35	M	46	9	6	6	5
R36	F	44	9	9	7	3
R37	F	42	8	5	8	4
R38	M	45	8	7	5	4
R39	F	39	7	7	6	5
R40	M	40	7	8	4	3

☐ once a week,
☐ once a month.

Q8. How often do you encounter issues related to mobile applications usability?

☐ very rarely (once a year),
☐ occasionally (once a month),
☐ moderately (once a week),
☐ frequently (several times a week),
☐ very often (several times a day).

Q9. How important is using mobile applications in your professional life?

☐ very high,
☐ high,
☐ moderate,
☐ low,
☐ very low,
☐ not applicable (I do not work).

Q10. How important is using mobile applications in your private life?

☐ very high,
☐ high,
☐ moderate,
☐ low,
☐ very low.

Q11. How do you evaluate your skills with regard to using mobile applications?

☐ very high,
☐ high,
☐ moderate,
☐ low,
☐ very low.

APPENDIX B
POST-TESTING QUESTIONNAIRE

We asked the users to evaluate the application effectiveness, efficiency, and satisfaction by using the following post-test

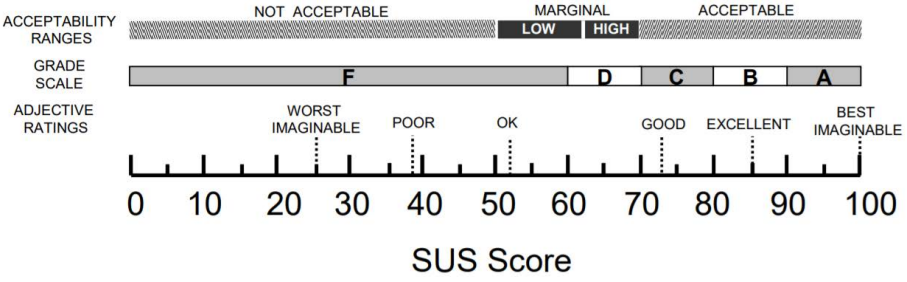

FIGURE 5. Adjective ratings, acceptability scores, and school grading scales, in relation to the average SUS score.

TABLE 12. Collected data from post-testing questionnaire regarding the perceived effectiveness and perceived efficiency.

ID	Sex	EFFE2	EFFE3	EFFE4	EFFE5	EFFI2	EFFI3	EFFI4	EFFI5	EFFI6
R1	M	6	7	7	7	9	8	8	7	7
R2	M	5	6	6	6	8	7	7	6	7
R3	M	6	7	7	7	7	8	8	7	8
R4	F	5	7	7	7	8	9	7	7	7
R5	F	6	6	7	6	7	7	7	7	7
R6	M	5	7	6	7	6	6	8	8	8
R7	F	6	6	7	6	7	7	8	9	8
R8	M	5	5	6	7	6	7	8	8	7
R9	F	6	6	6	7	8	8	7	7	7
R10	M	5	4	5	6	7	9	8	7	7
R11	F	6	5	6	7	7	8	7	8	7
R12	M	5	7	7	6	8	9	7	7	6
R13	M	6	6	7	7	8	8	7	6	8
R14	F	5	6	7	6	9	7	8	7	9
R15	M	6	7	6	7	9	8	7	9	8
R16	M	5	7	7	6	9	9	7	8	8
R17	M	5	7	7	7	9	8	7	8	7
R18	M	6	5	6	6	9	7	7	8	8
R19	F	6	6	7	7	7	7	8	7	8
R20	M	5	7	7	6	8	8	8	8	8
R21	M	6	6	6	7	6	7	7	7	8
R22	F	6	5	7	6	7	7	8	8	9
R23	M	5	6	6	7	8	8	8	7	8
R24	M	6	5	5	6	9	9	7	7	7
R25	M	5	5	5	7	7	8	8	8	7
R26	M	6	6	5	6	8	7	8	9	6
R27	F	7	5	5	7	7	9	7	7	9
R28	F	6	5	6	7	8	7	7	8	9
R29	M	5	5	6	6	9	7	7	8	8
R30	F	6	7	7	7	8	7	8	7	7
R31	M	5	7	7	6	7	7	8	8	7
R32	F	6	7	6	7	8	8	7	8	8
R33	M	6	5	7	6	9	9	8	9	8
R34	F	5	5	7	7	7	7	9	8	9
R35	M	6	5	6	6	6	7	7	7	8
R36	F	7	6	7	7	7	8	7	8	7
R37	F	6	5	6	6	8	9	8	7	7
R38	M	5	5	7	7	8	7	7	8	7
R39	F	4	5	6	6	9	8	7	7	8
R40	M	5	6	7	7	7	8	8	8	9

questionnaire. The short instructions below provide brief information regarding each of the attributes.

B. PERCEIVED EFFECTIVENESS MEASUREMENT TOOL

Please evaluate the following four statements by using the below 7-point Likert scale, ranging from: (1) absolutely inappropriate, (2) inappropriate, (3) slightly inappropriate, (4) neutral, (5) slightly appropriate, (6) appropriate to (7) absolutely appropriate.

- ☐ Total number of steps required to complete a task.
- ☐ Total number of taps related to app usage.
- ☐ Total number of taps unrelated to app usage.
- ☐ Total number of times the back button was used.

TABLE 13. Collected data from the SUS (post-testing) questionnaire, along with estimated SUS scores and assigned scales and their verbal interpretations.

ID	Sex	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SUS score	Scale	Adj
R1	M	4	2	4	1	5	1	4	1	4	3	82,5	B	good
R2	M	4	1	5	1	5	1	5	2	4	2	90	A	excellent
R3	M	5	2	4	1	5	1	4	1	5	3	87,5	B	excellent
R4	F	5	1	4	1	4	1	4	1	4	2	87,5	B	excellent
R5	F	5	1	5	2	5	2	5	2	5	1	92,5	A	excellent
R6	M	4	2	4	1	5	1	5	1	4	2	87,5	B	excellent
R7	F	5	2	5	2	4	1	5	1	5	3	87,5	B	excellent
R8	M	4	1	4	2	5	2	4	1	4	2	82,5	B	good
R9	F	5	2	5	1	5	1	5	2	5	1	95	A	excellent
R10	M	4	1	4	2	4	1	4	1	4	2	82,5	B	good
R11	F	5	1	5	2	5	2	5	2	5	3	87,5	B	excellent
R12	M	4	1	5	1	4	1	4	1	4	2	87,5	B	excellent
R13	M	5	2	5	1	5	2	5	1	4	3	87,5	B	excellent
R14	F	4	2	4	1	4	1	4	1	5	2	85	B	excellent
R15	M	5	1	4	1	5	2	5	1	5	1	95	A	excellent
R16	M	5	1	4	1	4	1	4	1	5	2	90	A	excellent
R17	M	5	1	5	1	5	2	5	2	4	3	87,5	B	excellent
R18	M	4	2	5	1	4	1	4	1	5	2	87,5	B	excellent
R19	F	5	1	5	1	5	2	5	1	4	1	95	A	excellent
R20	M	5	1	5	1	4	1	4	1	5	2	92,5	A	excellent
R21	M	5	2	4	1	5	2	5	1	4	1	90	A	excellent
R22	F	5	2	4	1	5	1	4	2	4	2	85	B	excellent
R23	M	4	1	4	1	4	2	5	1	5	3	85	B	excellent
R24	M	4	2	5	1	5	1	4	1	5	2	90	A	excellent
R25	M	5	2	5	1	4	2	4	1	5	1	90	A	excellent
R26	M	5	1	5	2	5	1	5	2	4	2	90	A	excellent
R27	F	5	2	4	1	5	2	4	1	5	3	85	B	excellent
R28	F	4	2	5	1	4	1	4	1	4	2	85	B	excellent
R29	M	5	2	4	1	5	2	5	1	5	1	92,5	A	excellent
R30	F	4	1	5	2	5	1	5	1	5	2	92,5	A	excellent
R31	M	5	1	4	1	4	2	5	2	5	1	90	A	excellent
R32	F	5	2	5	1	5	1	4	1	4	2	90	A	excellent
R33	M	4	1	4	1	4	2	5	1	4	3	82,5	B	good
R34	F	5	2	5	1	5	1	4	1	4	2	90	A	excellent
R35	M	4	1	5	2	5	2	5	2	5	2	87,5	B	excellent
R36	F	5	2	5	1	4	1	4	1	5	1	92,5	A	excellent
R37	F	4	1	4	1	5	1	5	1	4	2	90	A	excellent
R38	M	4	2	4	2	5	1	4	2	5	3	80	B	good
R39	F	5	1	4	1	4	1	5	1	4	2	90	A	excellent
R40	M	5	2	4	1	4	1	4	2	5	1	87,5	B	excellent

TABLE 14. Summary of the Task1 statistics of the observed effectiveness indicators.

Statistics / Gender / Indicators		EFFE2	EFFE3	EFFE4	EFFE5
mean (M)	Females	6.31	22.90	0.5	0.563
	Males	6.25	22.50	0.292	0.375
standard deviation (SD)	Females	0.48	2.28	0.516	0.892
	Males	0.53	2.32	0.464	0.77
mean difference (MD)		0.063	0.333	0.208	0.188
Shapiro-Wilk's p-value	Females	< 0.001	0.190	< 0.001	< 0.001
	Males	< 0.001	0.058	< 0.001	< 0.001
Levene's test (F-test)	F-value	0.130	0.208	3.190	0.717
	p-value	0.721	0.651	0.082	0.402
Student's test	t-value	0.379	0.448	1.33	0.708
	p-value	0.707	0.657	0.192	0.483
Effect size (Cohen's d)		0.122	0.145	0.429	0.229
Mann-Whitney U test	U value	175	180	152	168
	p-value	0.533	0.748	0.194	0.421

C. PERCEIVED EFFICIENCY MEASUREMENT TOOL

Please evaluate the following five statements, by using the 9-point Likert scale, with fixed endpoints, running from: (1) very low, (3) low, (5) moderate, (7) high, (9) very high, with four middle values.

- ☐ Duration of the application starting*.
- ☐ Duration of the application closing*.
- ☐ Duration of content loading*.
- ☐ Application performance continuity.
- ☐ Duration of the application response to the performed action*.

In the case of the statements marked with an asterisk (*), the reverse scale was used.

TABLE 15. Calculated statistics for the perceived effectiveness indicators.

Statistics / Gender / Indicators		EFFE2	EFFE3	EFFE4	EFFE5
mean (*M*)	Females	5.81	5.75	6.5	6.63
	Males	5.42	5.96	6.29	6.5
standard deviation (*SD*)	Females	0.75	0.78	0.623	0.5
	Males	0.50	0.96	0.751	0.511
mean difference (*MD*)		0.396	-0.208	0.208	0.125
Shapiro-Wilk's *p*-value	Females	0.004	0.002	< 0.001	< 0.001
	Males	< 0.001	< 0.001	< 0.001	< 0.001
Levene's test (F-test)	F-value	0.182	1.02	0.728	1.52
	p-value	0.672	0.32	0.399	0.225
Student's test	t-value	2.000	-0.727	0.914	0.765
	p-value	0.053	0.472	0.367	0.449
Effect size (Cohen's *d*)		0.646	-0.235	0.295	0.247
Mann-Whitney U test	U value	125	166	165	168
	p-value	0.04	0.448	0.411	0.452

TABLE 16. Summary of task completion times (EFFI1) and corresponding statistics (in seconds).

Statistics / Gender / Task		Task1	Task2	Task3	Task4	Task5
mean (*M*)	Females	41.94	8.50	7.31	6.31	4.13
	Males	42.3	8.83	6.92	6.17	4.38
standard deviation (*SD*)	Females	2.54	1.37	1.08	1.2	0.957
	Males	3.52	1.34	1.1	1.09	1.06
mean difference (*MD*)		-0.313	-0.333	0.396	0.146	-0.25
Shapiro-Wilk *p*-value	Females	0.685	0.610	0.19	0.218	0.026
	Males	0.611	0.062	0.069	0.055	0.007
Levene's test (F-test)	F-value	2.590	0.145	0.0001	0.249	0.576
	p-value	0.116	0.705	0.990	0.621	0.453
Student's test (T-test)	t-value	0.306	-0.765	1.12	0.399	-0.761
	p-value	0.762	0.449	0.268	0.692	0.451
Effect size (Cohen's *d*)		0.099	-0.247	0.363	0.129	-0.246

TABLE 17. Calculated statistics for the perceived efficiency indicators.

Statistics / Gender / Indicators		EFFI2	EFFI3	EFFI4	EFFI5	EFFI6
mean (*M*)	Females	7.63	7.69	7.5	7.5	7.88
	Males	7.79	7.75	7.5	7.63	7.5
standard deviation (*SD*)	Females	0.72	0.79	0.632	0.632	0.885
	Males	1.10	0.85	0.511	0.824	0.722
mean difference (*MD*)		-0.167	-0.063	0	-0.125	0.38
Shapiro-Wilk's *p*-value	Females	0.001	0.001	< 0.001	< 0.001	0.001
	Males	0.002	0.004	< 0.001	0.006	< 0.001
Levene's test (F-test)	F-value	3.770	0.025	1.520	1.09	1.44
	p-value	0.060	0.875	0.225	0.303	0.237
Student's test	t-value	-0.533	-0.234	0	-0.513	1.47
	p-value	0.597	0.816	1	0.611	0.15
Effect size (Cohen's *d*)		-0.172	-0.076	0	-0.166	0.474
Mann-Whitney U test	U value	167	181	186	170	153
	p-value	0.48	0.756	0.862	0.509	0.254

D. SATISFACTION MEASUREMENT TOOL

Introduced by John Brooke in 1996 [30], the System Usability Scale (SUS) is a low-cost method employed by researchers to assess usability. By design, the SUS forms a unidimensional measure of perceived usability. It should be noted here that, in this study, similarly to others [50], its use concerns the measurement of the notion of satisfaction.

Scores were calculated according to Brooke's guidelines in the following way. The SUS questionnaire consists of ten items (see Table 5). Each item's score contribution ranges from 0 to 4. For the positive items (odd numbers: 1, 3, 5, 7, and 9), the score contribution is the scale position minus 1, while for the negative items (even numbers: 2, 4, 6, 8, and 10), the score contribution is 5 minus the scale position. The final sum of all scores is then multiplied by 2.5 to get the final satisfaction value. Brooke used a 5-point Likert scale to measure the level of agreement with each of the items, provided with response choices ranging from (1) "*strongly disagree*" to (5) "*strongly agree*." Ultimately, SUS scores have a range of 0 to 100.

The SUS scale is intuitive to understand, but raises many questions about its meaning in an absolute sense.

TABLE 18. Calculated statistics for the perceived satisfaction (SUS scores).

Statistics / Gender		Value
mean (*M*)	Females	89.40
	Males	87.7
standard deviation (*SD*)	Females	3.48
	Males	3.68
mean difference (*MD*)		1.670
Shapiro-Wilk's *p*-value	Females	0.100
	Males	0.086
Levene's test (F-test)	F-value	0.045
	p-value	0.834
Student's test	t-value	1.430
	p-value	0.160
Effect size (Cohen's *d*)		0.463

To overcome this limitation, Bangor et al. [39] assigned grades as a function of SUS scores ranging from **F** (*worst imaginable*) to **A** (*best imaginable*), as is shown in Figure 5.

APPENDIX C
COLLECTED DATA

See Tables 6–13.

APPENDIX D
CALCULATED STATISTICS

See Tables 14–18.

REFERENCES

[1] A. M. Research. (2019). *Mobile Application Market by Marketplace and App Category: Global Opportunity Analysis and Industry Forecast, 2019–2026*. [Online]. Available: https://www.alliedmarketresearch.com/press-release/mobile-application-market.html

[2] Grand View Research. (2021). *Mobile Application Market Size, Share & Trends Analysis Report by Store Type*. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/mobile-application-market

[3] R. Nelson. (2020). *Consumer Spending in Mobile Apps Grew 17% in 2019 to Exceed $83 Billion Globally*. [Online]. Available: https://sensortower.com/blog/app-revenue-and-downloads-2019

[4] N. Gilbert. (2021). *70 App Statistics You Can't Ignore: 2021–2022 Market Share & Data Analysis*. [Online]. Available: https://financesonline.com/app-statistics/

[5] MobieleApps.com. (2021). *How Many Apps Are There Globally? (2021 Facts and Statistics)*. [Online]. Available: https://www.mobileapps.com/blog/how-many-apps-are-there

[6] Statista. (2021). *Mobile Internet Usage Worldwide—Statistics & Facts*. [Online]. Available: https://www.statista.com/topics/779/mobile-internet/

[7] Statista. (2022). *Number of Smartphone Subscriptions Worldwide From 2016 to 2027*. [Online]. Available: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[8] Statista. (2022). *Time Spent Per Day With Mobile Non-Voice Media in the United States From 2019 to 2023, by Type*. [Online]. Available: https://www.statista.com/statistics/469983/time-spent-mobile-media-type-usa/

[9] C. Coursaris and D. Kim, "A meta-analytical review of empirical mobile usability studies," *J. Usability Stud.*, vol. 6, no. 3, pp. 117–171, 2011.

[10] C.-L. Hung, J. C.-L. Chou, and C.-M. Ding, "Enhancing mobile satisfaction through integration of usability and flow," *Eng. Manage. Res.*, vol. 1, no. 1, p. 44, Apr. 2012.

[11] H. Khalid, E. Shihab, M. Nagappan, and A. E. Hassan, "What do mobile app users complain about?" *IEEE Softw.*, vol. 32, no. 3, pp. 70–77, May 2015.

[12] J. Dąbrowski et al., "Analysing app reviews for software engineering: A systematic literature review," *Empirical Softw. Eng.*, vol. 27, 2022, Art. no. 43, doi: 10.1007/s10664-021-10065-7.

[13] C. Iacob, V. Veerappa, and R. Harrison, "What are you complaining about?: A study of online reviews of mobile applications," in *Proc. Electron. Workshops Comput.*, Sep. 2013, pp. 1–6.

[14] M. Tavakoli, L. Zhao, A. Heydari, and G. Nenadić, "Extracting useful software development information from mobile application reviews: A survey of intelligent mining techniques and tools," *Expert Syst. Appl.*, vol. 113, pp. 186–199, Dec. 2018.

[15] E. G. Nilsson, "Design guidelines for mobile applications," SINTEF ICT, Trondheim, Norway, SINTEF Rep. STF90 A06003, Jun.2008.

[16] G. Ortiz, A. García-De-Prado, J. Berrocal, and J. Hernández, "Improving resource consumption in Context- aware mobile applications through alternative architectural styles," *IEEE Access*, vol. 7, pp. 65228–65250, 2019.

[17] UX Design Institute. (2020). *Industry Reports. Women in UX: An Industry Insight*. [Online]. Available: https://www.uxdesigninstitute.com/blog/women-in-ux-an-industry-insight/

[18] United Nations. (2022). *Goal 5: Achieve Gender Equality and Empower All Women and Girls*. [Online]. Available: https://www.un.org/sustainabledevelopment/gender-equality/

[19] SurveyMonkey Intelligence. (2016). *The 40 Top Apps for Men, According to App Demographic Data*. [Online]. Available: https://medium.com/sm_app_intel/the-40-top-apps-for-men-according-to-app-demographic-data-9bc77b9e88a1

[20] SurveyMonkey Intelligence. (2016). *The 40 top Apps for Women: Revealing the Apps That Ladies Love*. [Online]. Available: https://medium.com/sm_app_intel/the-40-top-apps-for-women-revealing-the-apps-that-ladies-love-8dd78a70247b

[21] D. Střelák, F. Škola, and F. Liarokapis, "Examining user experiences in a mobile augmented reality tourist guide," in *Proc. 9th ACM Int. Conf. Pervasive Technol. Rel. Assistive Environ.*, Jun. 2016, pp. 1–8.

[22] E. O. Mkpojiogu and N. L. Hashim, "The impact of users' age, gender, education, and experience on their satisfaction perception of m-banking app's usability," Malaysian Acad. Library, Malaysia, Tech. Rep., 2017.

[23] K. Oyibo and J. Vassileva, "The interplay of aesthetics, usability and credibility in mobile website design and the effect of gender," *SBC J. Interact. Syst.*, vol. 8, no. 2, pp. 4–19, 2017.

[24] K. C. Lim, A. Selamat, N. A. M. Ghani, M. H. M. Zabil, R. A. Alias, and F. Puteh, "Pre-processing of gender-based comparative usability performance data in mobile augmented reality English language teaching," in *Proc. IEEE Conf. e-Learn., e-Manag. e-Services (IC e)*, Nov. 2017, pp. 102–107.

[25] E. Ibili and M. Billinghurst, "Assessing the relationship between cognitive load and the usability of a mobile augmented reality tutorial system: A study of gender effects," *Int. J. Assessment Tools Educ.*, vol. 6, no. 3, pp. 378–395, Jul. 2019.

[26] S. Li and C.-H. Chen, "Effects of progress information and gender on time perception, usability, and emotion of the mobile wayfinding application interface," *J. Sci. Des.*, vol. 4, no. 2, pp. 211–218, 2020.

[27] International Organization for Standardization, *Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts*, Standard ISO 9241-11:2018(en), 2018. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

[28] P. Weichbroth, "Usability of mobile applications: A systematic literature study," *IEEE Access*, vol. 8, pp. 55563–55577, 2020.

[29] P. Weichbroth, "A mixed-methods measurement and evaluation methodology for mobile application usability studies," in *Proc. Commun. Papers Federated Conf. Comput. Sci. Inf. Syst.*, Sep. 2019, pp. 101–106.

[30] J. Brooke, "SUS—A quick and dirty usability," *Usability Eval. Ind.*, vol. 189, pp. 4–7, Dec. 1996.

[31] A. Kaya, R. Ozturk, and C. A. Gumussoy, "Usability measurement of mobile applications with system usability scale (SUS)," in *Proc. Ind. Eng. Big Data Era*. Cham, Switzerland: Springer, 2019, pp. 389–400.

[32] M. K. Felton, *Metacognitive Reflection and Strategy Development in Argumentive Discourse*. New York, NY, USA: Columbia Univ., 1999.

[33] M. Hertzum, P. Borlund, and K. B. Kristoffersen, "What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions," *Int. J. Hum.-Comput. Interact.*, vol. 31, no. 9, pp. 557–570, Sep. 2015.

[34] Nielsen Norman Group. (2021). *Remote Usability Tests: Moderated and Unmoderated*. Accessed: Jul. 26, 2021. [Online]. Available: https://www.nngroup.com/articles/remote-usability-tests/

[35] A. Woolrych, K. Hornbæk, E. Frøkjær, and G. Cockton, "Ingredients and meals rather than recipes: A proposal for research that does not treat usability evaluation methods as indivisible wholes," *Int. J. Hum.-Comput. Interact.*, vol. 27, no. 10, pp. 940–970, Oct. 2011.

[36] Jamovi. (2021). *About*. Accessed: Mar. 18, 2021. [Online]. Available: https://www.jamovi.org/about.html
[37] J. Cohen, *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, NJ, USA: Laurence Erlbaum Associates, 1988.
[38] S. S. Sawilowsky, "New effect size rules of thumb," *J. Modern Appl. Stat. Methods*, vol. 8, no. 2, p. 26, 2009.
[39] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, May 2009.
[40] Z. Huang and J. Mou, "Gender differences in user perception of usability and performance of online travel agency websites," *Technol. Soc.*, vol. 66, Aug. 2021, Art. no. 101671.
[41] P.-C. Lin and S.-I. Chen, "The effects of gender differences on the usability of automotive on-board navigation systems—A comparison of 2D and 3D display," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 19, pp. 40–51, Jul. 2013.
[42] E. Ozdemir and S. Kilic, "Young consumers' perspectives of website visualization: A gender perspective," *Bus. Econ. Res. J.*, vol. 2, no. 2, pp. 41–60, 2011.
[43] J. Hou and K. Elliott, "Gender differences in online auctions," *Electron. Commerce Res. Appl.*, vol. 17, pp. 123–133, May 2016.
[44] B. Hasan, "Exploring gender differences in online shopping attitude," *Comput. Hum. Behav.*, vol. 26, no. 4, pp. 597–601, Jul. 2010.
[45] C. C. Hsu, "Comparison of gender differences in young people's blog interface preferences and designs," *Displays*, vol. 33, no. 3, pp. 119–128, Jul. 2012.
[46] C. Wang, H. Qu, and M. K. Hsu, "Toward an integrated model of tourist expectation formation and gender difference," *Tourism Manage.*, vol. 54, pp. 58–71, Jun. 2016.
[47] W. O. Peake, W. C. McDowell, M. L. Harris, and P. E. Davis, "Can women entrepreneurs plan to prosper? Exploring the role of gender as a moderator of the planning-performance relationship," in *Inside Mind Entrepreneur*. Cham, Switzerland: Springer, 2018, pp. 121–133.
[48] M. P. Matud, M. López-Curbelo, and D. Fortes, "Gender and psychological well-being," *Int. J. Environ. Res. Public Health*, vol. 16, no. 19, p. 3531, 2019.
[49] M. S. Rahman, "The advantages and disadvantages of using qualitative and quantitative approaches and methods in language 'testing and assessment' research: A literature review," Tech. Rep., 2020. [Online]. Available: https://files.eric.ed.gov/fulltext/EJ1120221.pdf
[50] N. Harrati, I. Bouchrika, A. Tari, and A. Ladjailia, "Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis," *Comput. Hum. Behav.*, vol. 61, pp. 463–471, Aug. 2016.

PAWEŁ WEICHBROTH (Member, IEEE) received the M.A. degree in statistics from the University of Gdańsk, Poland, in 2003, and the Ph.D. degree in artificial intelligence from the Katowice University of Economics, Poland, in 2014.

He is currently an Assistant Professor with the Department of Software Engineering, Gdańsk University of Technology. Moreover, for over 20 years, he has worked as a Business Consultant and IT Lecturer. Since 2018, he has been an Expert for the Ministry of Digital Affairs in a project for the development of public digital services. His main research interests include software quality, machine learning, and knowledge management. In this regard, he has authored over 40 research papers as journal articles, conference papers, and book chapters. He has been a member of the Scientific Community of Business Informatics, and several international conference program committees, actively acting as a reviewer and as an organizer.