applied sciences

MDPI

Article

Usability Testing of Mobile Applications: A Methodological Framework

Paweł Weichbroth

Department of Software Engineering, Faculty of Electronics, Gdansk University of Technology, 80-233 Gdansk, Poland; pawel.weichbroth@pg.edu.pl

Abstract: Less than five percent of all mobile applications have become successful throughout 2023. The success of a new mobile application depends on a variety of factors ranging from business understanding, customer value, and perceived quality of use. In this sense, the topic of usability testing of mobile applications is relevant from the point of view of user satisfaction and acceptance. However, the current knowledge seems to be fragmented, scattered across many papers and reports, and sometimes poorly documented. This paper attempts to fill this gap by investigating the current state of knowledge by reviewing the previous literature relevant to the research topic and developing a unified view. In particular, the methodological framework is outlined and discussed, including the discourse on settings for laboratory and field studies, data collection techniques, experimental designs for mobile usability testing, and a generic research framework. Therefore, the paper contributes to both the theory and practice of human–computer interaction by providing methodological foundations for usability testing of mobile applications, paving the way for further studies in this area. Moreover, the paper provides a better understanding of the related topics, in particular shedding light on methodological foundations, key concepts, challenges, and issues, equipping readers with a comprehensive knowledge base to navigate and contribute to the advancement of the field of mobile usability.

Keywords: mobile usability; testing; methodology; framework

Citation: Weichbroth, P. Usability Testing of Mobile Applications: A Methodological Framework. *Appl. Sci.* **2024**, *14*, 1792. https://doi.org/10.3390/app14051792

Academic Editor: Andrea Prati

Received: 17 December 2023
Revised: 5 February 2024
Accepted: 20 February 2024
Published: 22 February 2024

1. Introduction

By the end of 2023, the number of mobile phone users, including both smart and feature phones, will reach 7.33 billion, which stands for 91.21 percent of the world's population [1]. What do mobile phone users spend most of their time doing? In April 2022, the average amount of time an American user spends on the phone each day, not counting calls, has increased to a total of 4 h and 30 min [2]. Interestingly, the average user checks their phone 344 times a day. That is once every 4 min [3]. Another study found that, in the US, the revenue from smartphone sales increased between 2013 and 2022, peaking at more than USD 102 billion in 2022 [4].

Despite the ubiquity of smartphones, only about 5 percent of mobile applications are successful in the marketplace [5]. As practice shows, around 80–90 percent of the applications published in the app stores are abandoned after just a single use [6]. From those remaining, about 77 percent lose their daily active users within the first three days after installation [7]. Moreover, according to Mobile Commerce Daily, about 45 percent of users dislike their mobile applications [8].

For a variety of reasons, the majority fall into the usability domain [9]. For instance, when considering mobile commerce, four factors, convenience, ease of use, trust, and ubiquity, were identified as the most important [10]. In fact, usability is often pointed out as one of the success factors for the adoption of mobile applications by users [11,12]. From the business perspective, poor usability can reduce employee productivity and undermine the overall value of a mobile enterprise solution [13].

From the user perspective, usability might be understood as the set of choices leading to accomplishing one or more specific tasks efficiently, effectively, and with minimum errors [14]. In other words, building successful software means reaching beyond codes and algorithms and into a genuine understanding of what your users do and how they do it. The benefits of usable applications concern reduced costs of training, support and service, as well as increased user productivity and satisfaction and application maintainability [15].

Obviously, the need for usability testing is nothing new as mobile software vendors are interested in whether or not their products are usable [16]. As mobile phones have rapidly evolved from simple communication devices to multifunctional multimedia systems [17], the need for effective usability testing has become paramount. Despite the growth in mobile human–computer interaction research [18], there is a research gap in comprehensive usability testing frameworks tailored to the ever-evolving functionalities of modern smartphones [19], along with the growing expectations and requirements of their users [20].

Given the wide range of techniques, methods, and frameworks that have already been adapted to mobile usability testing from the computer and social sciences, our goal is to generalize across this body of literature (rather than provide an exhaustive list of them) and develop a unified methodological framework. In addition, we attempt to address some of the key challenges of mobile usability testing by drawing on the latest research by synthesizing the wealth of existing knowledge into a cohesive and flexible approach that can serve as a guide for researchers and practitioners.

For this purpose, we used both well-known databases for peer-reviewed literature and books (Google Scholar, ACM, IEEE, and Scopus) as well as gray literature using the Google search engine. To identify relevant documents, we relied on keywords extracted from the name of the search topic of current interest, such as “testing framework”, “mobile”, “usability”, “testing methods”, as well as their combinations by adding the logical conjunction operator. To this end, we followed the guidelines and recommendations developed by Whittemore and Knafl [21].

The remainder of the paper is organized into four sections. In Section 2, related studies are briefly reviewed. In Section 3, the theoretical background is presented. In Section 4, the methodological framework is outlined, followed by Section 5, which presents its use cases. In Section 6, a comprehensive discussion is carried out, followed by Section 7, which concludes the paper.

2. Related Work

To date, many studies have been conducted on usability in the context of mobile applications. To the best of our knowledge, the majority of the research has focused on the evaluation of specific applications, adopting and adapting methods and tools that are well-established in desktop usability research. Thus, few studies have attempted to provide an in-depth analysis of the existing methods, tools, and approaches regarding mobile usability testing. However, there are a few worth mentioning, and these are discussed below.

Zhang and Adipat [22] developed a generic framework for conducting usability testing for mobile applications based on a comprehensive review and discussion of research questions, methodologies, and usability attributes. The authors’ framework was designed to guide researchers in selecting valid testing methods, tools, and data collection methods that correspond to specific usability attributes. In addition, the authors identified six key challenges, including mobile context, connectivity, small screen size, multiple display resolutions, limited processing capability and power, and multimodal input methods.

Ji et al. [23] introduced and developed a task-based usability checklist based on heuristic evaluation in terms of mobile phone user interface (UI). To address the challenges of usability evaluation, a hierarchical structure of UI design elements and usability principles related to mobile phones was developed and then used to develop the checklist. The developed usability checklist is mainly based on heuristic evaluation methods, which are the most popular usability evaluation methods. The authors argued that, while the

effectiveness of heuristic evaluation is closely related to the importance of selecting usability guidelines, the corresponding 21 usability principles were developed, which are crucial in mobile phone UI design. Interestingly, the authors suggested that certain usability features, initially perceived as attractive or novel, eventually become essential to users. For example, features such as built-in cameras in mobile phones, once considered luxuries, have become common expectations.

Au et al. [13] developed the Handheld device User Interface Analysis (HUIA) testing framework. The rationale behind such a tool concerns effective usability testing, improved accuracy, precision, and flexibility, as well as reduced resource requirements such as time, personnel, equipment, and cost. Automating mobile usability testing can improve testing efficiency and ease its integration into the development process. While this paper demonstrates an effective tool, it says little about the theoretical aspects of usability testing.

Heo et al. [24] proposed a hierarchical model consisting of four levels of abstraction:

1. Mobile phone usability level. The level indicates what we ultimately want to evaluate. As an emergent concept, usability cannot be measured directly or precisely. Instead, it could be indirectly indicated as the sum of some usability factors underlying the concept of usability.
2. Usability indicator level. Five usability indicators are relevant to the usability of mobile phones, including visual support of task goals, support of cognitive interaction, support of efficient interaction, functional support of user needs, and ergonomic support.
3. Usability criteria level. This level identifies several usability factors that can be directly measured using different methods.
4. Usability property level. This level represents the actual states or behaviors of several interface features of a mobile phone, providing an actual usability value to the criteria level.

Since there are goal–means relationships between adjacent levels, accordingly to the authors, a variety of usability issues can be interpreted in a comprehensive manner as well as diagnostically. The model supports two different types of evaluation approaches, namely task-based and interface-based, supplemented by four sets of checklists.

Husain and Kutar [25] reviewed the current practices of measuring usability and developed the guidelines to guide mobile application developers, which later served as the basis for developing the GQM model. The model is based on three measures, including effectiveness, efficiency, and satisfaction, along with corresponding goals and guidelines. The presented model seems to be of great value for usability practitioners, but little is discussed about methodological aspects of mobile usability testing.

JongWook et al. [26] presented methods and tools for detecting mobile usability issues through testing, expecting that users who interact with mobile applications in different ways would encounter a variety of usability problems. Based on this idea, the authors proposed a method to determine the tasks that contain usability issues by measuring the similarity of user behavior, with the goal of automatically detecting usability problems by tracking and analyzing user behavior. The empirical results showed that the developed method seems to be useful for testing large systems that require a significant number of tasks and could benefit software developers who are interested in performing usability testing but have little experience in this area.

In conclusion, following a literature review of the usability studies dedicated to mobile application testing, a majority focused on the design, development, and empirical evaluation of different methods and tools. However, in a few studies, the research attention was directed to revisiting the theoretical foundations that provide all the necessary means to organize and conduct a correct usability study that takes advantage of these methods and tools. In no way do we question the validity of research that presents practical and useful approaches, but theory seems essential for their effective development and implementation.

3. The Theory of Mobile Usability

3.1. Usability Conceptualization

In light of a recent study [27], the most widely accepted definition of usability is that provided in the ISO 9241-11 standard, which states that usability is "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness efficiency and satisfaction in a specified context of use" [28].

At this point, of an obvious nature, a question arises: how does one understand context? Context can be understood as a carrier of information about the environment, place, time, and situation in which an entity currently exists [29]. Here, an entity is a user who deliberately interacts with a mobile application. With the ubiquity of mobile devices (GPS devices) [30] and Internet connectivity (public Wi-Fi hotspots, home Wi-Fi, LTE 4G, and 5G) [31], the ability to incorporate this type of information is common and in many domains has become even an imperative to use [32]. In summary, context in mobile systems can be divided into three categories [33]:

- external, independent, and valid for all interested users (e.g., current weather, dangerous events, time, etc.);
- location, refers to information about the user's point of interest (e.g., traffic jams, road conditions, parking space, restaurant reviews, etc.);
- user-specific, related to the user's attributes, beliefs, activities, and interests (e.g., gender, age, nationality, religion, etc.).

Incorporating such context in mobile applications significantly enhances the quality of service in terms of perceived usefulness by making our everyday environments increasingly intelligent [34].

3.2. Usability Attributes Conceptualization

By definition, an attribute is a quality or feature regarded as a characteristic or inherent part of something [35]. Similarly to the notion of usability, attributes do not exist as such. On the contrary, they emerge from the physical interaction between the user and the mobile application. If now one takes into account the aforementioned usability definition, the question arises as to how to measure the extent of effectiveness, efficiency, and satisfaction. The answer is twofold: through user observation or by user survey.

That being said, an attribute can be classified as "observable" or as "perceived", respectively. While it is possible to change the type from the former to the latter, then the reverse operation is hardly achievable, or even impossible, due to human nature. For instance, very few users, if any, explicitly manifest satisfaction during or after using typical mobile applications. Nevertheless, there have been attempts to identify, measure, and evaluate numerous qualities with regard to both the user and the application, especially in domains such as games [36] or entertainment [37,38].

Let us now look at three attributes referred to in the ISO 9241-11 standard. It should be noted that, while effectiveness and efficiency are directly observable qualities, satisfaction is a "hidden" quality. Moreover, it is also possible to measure both effectiveness and efficiency through user survey. In short, Table 1 shows the 2-category classification of the ISO 9241-11 usability attributes.

Table 1. Usability attributes classification.

Attribute/Type	Observed	Perceived
Effectiveness	•	•
Efficiency	•	•
Satisfaction	Not applicable	•

Such a distinction has implications for the conceptualization of the usability attributes. Firstly, in the case of the observed category, the object of measurement is a user, or, more

precisely, the user's level of task performance. With this assumption, Table 2 shows the definitions of the observed usability attributes.

Table 2. The conceptualization of the observed usability attributes.

Attribute	Definition
Effectiveness	the ability of a user to complete a task in a given context [39]
Efficiency	the ability of a user to complete a task with speed and accuracy [40]
Satisfaction	Not applicable

Secondly, in the case of the second category, the object of measurement is a mobile application, in particular the user's perceived level of workload and application performance, as well as the self-reported level of satisfaction. The definitions of the perceived usability attributes are provided in Table 3.

Table 3. The conceptualization of the perceived usability attributes.

Attribute	Definition
Effectiveness	a user's perceived level of workload in a given context [41]
Efficiency	a user's perceived level of application performance (in terms of time) [42]
Satisfaction	a user's perceived level of comfort and pleasure [43]

In summary, the observed attributes can be interpreted in terms of the performance-based characteristics of the user, whereas the perceived attributes can be interpreted in terms of the user's perceptions of certain application characteristics, as well as their own feelings of comfort and task fulfilment.

It should also be noted that there are other commonly studied attributes that are considered latent variables. In this regard, the most frequent ones also concern [27] learnability, memorability, cognitive load, simplicity, and ease of use.

3.3. Usability Attributes Operationalization

By definition, operationalization is "the process by which a researcher defines how a concept is measured, observed, or manipulated within a particular study" [44]. More specifically, the researcher translates the conceptual variable of interest into a set of specific "measures" [45]. Note that, here, a measure is a noun and means a way of measuring with the units used for stating the particular property (e.g., size, weight, and time), whereas "measures of quantitative assessment commonly used for assessing, comparing, and tracking performance or production" are termed as metrics [46]. In other words, a metric is a quantifiable measure of the observed variable.

However, the other way to quantify variables is to use indicators. By definition, an indicator is "a quantitative or qualitative variable that provides reliable means to measure a particular phenomenon or attribute" [47]. Indicators are used to operationalize latent variables [48], in both reflective and formative measurement models [49]. In summary, for the sake of methodological clarity of the above terms "metric" and "indicator", only the former will be used for both observable and perceived attributes.

Drawing upon the usability attributes classification, now we can turn to operationalize them, which requires specification of the quantifiable metrics, along with corresponding measurement scales.

3.3.1. Observed Effectiveness

To quantify the observed effectiveness of a user in the context of the performed tasks, in total, five metrics are provided in Table 4 with assigned units and quantities.

Table 4. The observed effectiveness metrics.

Code	Metric	Unit and Quantity
EFFE1	rate of successful task completion	integer/amount
EFFE2	total number of steps required to complete a task	integer/amount
EFFE3	total number of taps related to app usage	integer/amount
EFFE4	total number of taps unrelated to app usage	integer/amount
EFFE5	total number of times that the back button was used	integer/amount

3.3.2. Observed Efficiency

By definition, efficiency is a quality that is measured by the amount of resources that are used by a mobile application to produce a given number of outputs. Now, thinking in terms of usability testing, the measured resource concerns the amount of time that a user needed to perform a particular task. Thus, the observed efficiency is measured by the completion time (EFFI1 metric) in units of time (commonly in seconds) with respect to each individual task, or much less often to a set of related tasks.

3.3.3. Perceived Effectiveness

It should be noted that observed and perceived effectiveness are measured by the same metrics except for the first one (EFFE1) since its submission to the respondent would imply a self-assessment of the rate of task completion. The following 7-point Likert scale can be used: absolutely inappropriate (1), inappropriate (2), slightly inappropriate (3), neutral (4), slightly appropriate (5), appropriate (6), and absolutely appropriate (7).

3.3.4. Perceived Efficiency

If we consider efficiency as an unobservable construct, the 7-point rating scale is also used to measure and rate the mobile application in this view. Table 5 shows the details of the perceived efficiency metrics.

Table 5. The perceived efficiency metrics [50].

Code	Metric	Scale
EFFI2	duration of the application starting *	7-point Likert scale
EFFI3	duration of the application closing *	
EFFI4	duration of content loading *	
EFFI5	duration of the application response to the performed actions *	
EFFI6	application performance continuity	

* the reverse scale.

Similarly, if efficiency is treated as an unobservable construct, the 7-point Likert rating scale can be used to measure and evaluate the mobile application in this perspective, starting from extremely low (1), very low (2), low (3), moderate (4), high (5), very high (6), to extremely high (7). Note that, for all metrics, expect the last one, a reverse scale must be used to estimate the perceived efficiency in order to preserve the correct interpretation of the collected data.

3.3.5. Perceived Satisfaction

In general, satisfaction is "a pleasant feeling you get when you get something you wanted or when you have done something you wanted to do" [51]. The perceived satisfaction construct (SATI) is composed of the three metrics validated in other usability studies. Table 6 provides a detailed description.

Table 6. The perceived satisfaction metrics.

Code	Metric	Scale
SATI1	I think I made the correct decision to use the *X* mobile application [52]	7-point Likert scale
SATI2	My experience using *X* mobile app has been satisfactory [53]	
SATI3	I am satisfied with the quality of *X* [54]	

X is the name of the mobile application being evaluated.

The following 7-point Likert scale can be used, starting with strongly disagree (1), disagree (2), somewhat disagree (3), neither agree nor disagree (4), somewhat agree (5), agree (6), and strongly agree (7).

4. Methodological Framework for Mobile Usability Testing

There is no consensus on the definition of usability testing. To this day, numerous attempts have been made in this respect. Let us look at just a few of these, which are well-accepted by the research community. So far, usability testing is

- "a technique used to evaluate a product by testing it on users" [55];
- "a technique for identifying difficulty that individuals may have using a product" [56];
- "a widely used technique to evaluate user performance and acceptance of products and systems" [57];
- "an essential skill for usability practitioners—professionals whose primary goal is to provide guidance to product developers for the purpose of improving the ease of use of their products" [58];
- "an evaluation method in which one or more representative users at a time perform tasks or describe their intentions under observation" [59];
- "a technique for ensuring that the intended users of a system can carry out the intended tasks efficiently, effectively and satisfactorily" [60] (borrowed from G. Gaffney [61])";
- "a systematic way of observing actual users trying out a product and collecting information about the specific ways in which the product is easy or difficult for them" [62].

As can be seen, the above definitions differ considerably. Firstly, some of them indicate that usability testing is 'just' an evaluation technique, while the last one refers to a systematic approach. Secondly, while some are general, others are precise by referring to the specific product (system) attributes. Thirdly, although it is not always explicitly acknowledged, a central role is played by the user, who interacts directly with a product (system) by carrying out a task or a set of tasks.

Having said that, and taking into account both the adopted definition of usability and the research context, usability testing is the process of evaluating a mobile application by specified users performing specified tasks to assess effectiveness, efficiency, and satisfaction in a specified context of use.

4.1. Research Settings for Usability Testing of Mobile Applications

First and foremost, there are two various approaches to usability evaluation, namely laboratory and field testing [63].

4.1.1. Laboratory Studies

The Usability Laboratory is an environment in which researchers are able to study and evaluate the usability of software products. One of the key requirements concerns comfortable conditions, which means the ability to provide sufficient physical space to accommodate a wide variety of study types, from those involving single users to those involving groups of collaborating users. The most favorable configuration is two separate rooms, with the first dedicated to a user and the second to a researcher.

Typically, a user testing laboratory is equipped with hardware equipment that includes, as a minimum configuration, the following items:

- a desk and chair, used by a user during application testing;
- document camera, a real-time image capture device, responsible for video recording;
- microphone, an external device that enables audio recording;
- video camera, a optical instrument that allows real-time observation of the user;
- a computer system unit (PC), optionally equipped with a keyboard, mouse, and external monitor, used to store session recordings; alternatively, a laptop computer may be used as an efficient substitute.

The observation room should generally be the same size or larger than the test laboratory, and should accommodate at least two observers at a time. A typical equipment configuration involves items such as

- monitor, used to view a user performing tasks;
- microphone, as a means of communication with a user;
- office equipment, necessary to ensure comfortable working conditions.

On the other hand, these two rooms can be integrated into a single laboratory shared by the user and the researcher. A studio of this kind is also suitable for carrying out usability tests and user surveys, provided that there is sufficient space and comfort for at least two persons (adults) at a time.

Two different evaluation approaches can be distinguished [64]:

- laboratory-based testing by using a computer-based mobile device emulator;
- laboratory-based testing by using a mobile device (smartphone or tablet).

By its very nature, usability testing places a strong emphasis on the solution of tasks or the achievement of goals by the specified users with the use of a product (system) in a given context [65]. It should be noted at this point that the role of context should not be underestimated [66]. Thus, using a mobile application emulator in the lab does not provide equivalent input capabilities to a real mobile device. In addition, the ability to emulate the context is limited. As a result, neither task performance nor context awareness appear to be measurable factors when testing mobile applications by using desktop simulators. Nevertheless, they could be reasonable choices for app prototyping.

For the reader interested in setting up a new laboratory, Schusteritsch et al. [67] provide an in-depth analysis of infrastructure and hardware equipment. In particular, the authors consider and describe the different designs and setups, as well as the factors affecting their effectiveness and efficiency.

4.1.2. Field Studies

By definition, a field study is understood as an investigation that is conducted in an environment that is not under the total control of the researcher [68]. Such a testing strategy assumes that the tasks are performed directly by the users in the field, in the workplace, or in any other non-laboratory location. Although there are no formal standards, it is widely accepted practice that users are not under any control. However, they are generally instructed as to the objectives of the study and the predefined tasks. One could also think of controlled laboratory research [69], that is, research that is carried out in an environment that is specifically designed for research [70].

The main advantage of field study is its generalizability to real-life contexts as it represents a greater variety of situations and environments that users experience in their natural environment. It is a powerful method for understanding the role of context in shaping user experience by providing a better understanding of subjective attitudes. A unique strength lies in its ability to reveal elements of users' experiences that we were previously unaware of [71].

On the other hand, embedded context includes possible factors that may affect the user while performing a task. These can be external, environmental, and personal influences. Let us consider three examples. In the case of external factors, a user is exposed to factors

beyond his or her control, such as blinding sunlight, deafening noise, heavy rain or snow, or strong wind. Environmental factors involve interaction with the environment. For example, a user standing on a bus, holding on to the handrail with one hand, and using a smartphone with the other hand, has limited capacity. Personal factors are related to motivation, pressure, and stress [72], which can have different causes and sources, related to both personal life and professional work.

From a practical point of view, field-based testing is carried out by means of a mobile device and a wireless camera, both attached to a portable stand, or by means of a screen recorder installed on a mobile device [73]. In addition, it is claimed that the process per se is time-consuming and complicates data collection [74].

4.1.3. Laboratory vs Field Studies

When comparing laboratory-based testing with field-based testing, both approaches have their own shortcomings. In the former, testing in different contexts seems to be limited by physical constraints, while, in the latter, there are several practical difficulties related to unfavorable weather conditions, pedestrian disturbance, and electronic equipment shortcomings [73]. However, since the user feedback comes from interacting with the system in a real environment, it provides more reliable and realistic information compared to a laboratory study [75].

While "testing out on the road" provides an opportunity to sample from a distributed user base [76], it has been rather rarely applied with respect to mobile applications [77]. Interestingly, the debate about the best site conditions and settings still seems to be an ongoing issue [78–80].

4.1.4. Self-User Usability Testing

Self-user usability testing involves mobile application testing in the "real world" by the users alone, without being instructed on how and when to use it. In other words, this type of testing is not intended to be driven by a set of instructions, recommendations, or task scenarios but rather to be performed in a completely free manner. Typically, such user testing emphasizes not only the usability attributes but in most cases broader perceptions and responses, termed as user experiences (UX).

This approach assumes that the user has a personal and unrestricted experience related to the use of a specific mobile application. A questionnaire is usually used to collect data, including questions on selected aspects of perceived usability. For this purpose, both desktop measurement tools, such as the Software Usability Scale (SUS) [81], as well as mobile-oriented instruments, such as Mobile Phone Usability Questionnaire (MPUQ) [82], are in common use.

Note that, in modern science, usability is often lumped under or related to user experience research, which encompasses all the effects that the use of a product has on the user, before, during, and after use [83], strongly influenced by the purpose of use and the context of use in general [84].

4.2. Data Collection Techniques

The literature review revealed that the empirical works concerning usability testing have taken advantage of both quantitative and qualitative research methods [85–87]. In a typical scenario, a usability testing session is a body of four integrated methods, namely:

1. Questionnaire.
2. Participant observation.
3. Thinking aloud.
4. Interview.

Drawing on the theory of academic research as well as recent literature addressing the issues related to mobile application usability research, each method is described in general terms and briefly placed in the context of the current paper.

4.2.1. Questionnaire

With regard to the first type of research method, it is important to distinguish between two terms that are sometimes used interchangeably, namely a questionnaire and a survey. In general, a questionnaire is a set of written questions used to collect information from a number of people [88], while a survey is an examination of opinions, behavior, etc., conducted by asking people questions [89]. As one can notice, the latter has a broader meaning. In addition, survey research can use qualitative research strategies (e.g., interviews with open-ended or closed-ended questions) [90], quantitative research strategies (e.g., questionnaires with numerically rated items) [91], or both strategies (a mixed methods approach) [92]. For the sake of clarity, survey will be referred to as a method, while both questionnaire and interview will be referred to as data collection techniques. In fact, there have been numerous surveys that have investigated mobile usability using different combinations of data collection techniques.

4.2.2. Participant Observation

Participant observation is a qualitative research method [93] where the researcher deliberately participates in the activities of an individual or group that is the subject of the research [94]. There are four different types of participation [95]:

- Passive occurs when the researcher is present but does not interact with people. At this level of participation, those being observed may not even be aware that they are being observed. By acting as a pure observer, a great deal of undisturbed information can be obtained [96].
- Moderate is when the observer is present at the scene of action. However, recognized as a researcher, the observer does not actively interact with those involved but may occasionally be asked to become involved. At this level of interaction, it is typical practice to use a structured observation framework. In some settings, moderate participation acts as a proxy until more active participation is possible.
- Active occurs when the researcher participates in almost everything that other people do with the aim of learning. In addition, a researcher proactively interacts with the participants (e.g., by talking to them and performing activities), thereby collecting all the necessary information.
- Complete is when the researcher is or becomes a member of the group that is being studied. To avoid disrupting normal activities, the role is usually hidden from the group [97].

In general, the methodology of participant observation is practiced as a form of case study, attempting to describe a phenomenon comprehensively and exhaustively in terms of a formulated research problem [98].

In usability studies, a typical setting for participant observation is passive. In this respect, a researcher acts as a moderator during a testing session. In addition, considering the presence of the moderator during the testing session, there are two types of usability testing approaches [99]:

- Moderated, requiring the moderator to be present either in person or on camera. In an in-person-moderated usability test, a moderator asks a participant to complete a series of tasks while observing and taking notes. So, both roles communicate in real-time.
- Unmoderated, which does not require the presence of a moderator. Participants perform application testing at their own pace, usually guided by a set of prompts.

It should be noted that both moderated and unmoderated sessions can be divided into local and remote studies. While the former involves the physical presence of a moderator, the latter can be conducted via the Internet or telephone. The unmoderated session is used when budget, time, and resources are limited [100]. In this line of thinking, such surveys are more efficient and less expensive; however, they are more suitable for larger pools of participants [101]. In addition, due to some participants' careless responding, the risk of the collection of unreliable data is also higher.

Usability testing sessions are typically recorded, allowing retrospective analysis that provides first-hand insight into interactions and behaviors [102]. In particular, detailed analysis allows task performance to be reconstructed. By extracting specific numerical values from a video recording, it is possible to calculate metrics of observed usability attributes, including both effectiveness and efficiency. In addition, the video recordings serve as a valuable reference [103], allowing interested parties to observe first-hand how users navigate through interfaces, interpret content, and respond to various features, ultimately facilitating data-driven decisionmaking and continuous improvement in the mobile application development process [104].

4.2.3. Thinking Aloud

Thinking aloud is the simultaneous verbalization of thoughts during the performance of a task [105]. It is interesting to note that in the literature one can come across two different names that are used as synonyms, namely "verbal protocol analysis" or "talking aloud" [106]. The basic assumption behind thinking aloud is that, when people talk aloud while performing a task, the verbal stream acts as a 'dump' of the contents of working memory [107]. According to Bernardini [108], under the right circumstances, i.e., verbally encoded information, no interference, no social interaction, and no instruction to analyze thoughts, it is assumed that such verbalization does not interfere with mental processes and provides a faithful account of the mental states occurring between them.

According to this view, the verbal stream can thus be viewed as a reflection of the cognitive processes used and, after analysis, provides the researcher with valuable ad hoc information about the user's perceptions and experiences during task performance. In this line of thinking, there are two types of thinking aloud usability tests: concurrent verbalization and retrospective [109]. Specifically, while the former requires participants to complete a task and narrate what is going through their minds, the latter relies on participants to report on their experiences after completing the task. Therefore, one can conclude that, by assumption, the well-known definition of thinking aloud actually refers to concurrent verbalization. Despite its value, analyzing think-aloud sessions can be tedious as they often involve evaluating all of a user's verbalization [110].

Moreover, different types of thinking aloud involve the verbalization of different types of information. Having said that, the modern literature usually tends to point to two different approaches [111]:

- Relaxed (or interactive) thinking aloud: a test user is asked to verbalize his or her thoughts by providing a running commentary on self-performed actions and, in moderated tests, is encouraged to self-reflect on current thoughts and actions [112].
- Classic thinking aloud: a test user is limited to verbalizing information that is used or has been used to solve a particular task. In addition, the interaction between researcher and user should be restricted to a simple reminder to think aloud if the user falls silent [113].

For obvious reasons, Ericsson and Simon's recommendations are expected to be followed in unmoderated test sessions. On the other hand, as usability testing practice shows, their guidelines are often relaxed in moderated test sessions, with the aim of eliciting a broader panorama of the user's thoughts and reflections [114].

The thinking aloud method has become popular in both practical usability evaluation and usability research, and is considered by many to be the most valuable usability evaluation method [115]. Nevertheless, while some researchers have found this method somewhat useful for mobile applications, especially for tasks of relatively low complexity [50], others have appreciated its effectiveness in identifying usability problems [116].

4.2.4. Interview

An interview is a qualitative research method that relies on asking people questions in order to collect primary data that are not available through other research methods. Typically, a researcher engages in direct conversation with individuals to gather information

about their attitudes, behaviors, experiences, opinions, or any other type of information. In the course of such a procedure, three different approaches can be applied [117]:

- An unstructured or non-directive interview is an interview in which the questions are not predetermined, i.e., the interviewer asks open-ended questions and relies on the freely given answers of the participants. This approach therefore offers both parties (interviewer and interviewee) flexibility and freedom in planning, conducting, and organizing the interview content and questions [118].
- A semi-structured interview combines a predetermined set of open-ended questions, which also encourage discussion, with the opportunity for the interviewer to explore particular issues or responses further. The rigidity of its structure can be varied depending on the purpose of the study and the research questions. The main advantages are that the semi-structured interview method has been found to be successful in enabling reciprocity between the interviewer and the participant [119], allowing the interviewer to improvise follow-up questions based on the participant's responses. The semi-structured format is the most commonly used interview technique in qualitative research [120].
- A structured interview is a systematic approach to interviewing in which you pose the same pre-defined questions to all participants in the same order and rate them using a standardized scoring system. In research, structured interviews are usually quantitative in nature. Structured interviews are easy to replicate because they use a fixed set of closed-ended questions that are easy to quantify and thus test for reliability [121]. However, this form of interview is not flexible; i.e., new questions cannot be asked off the cuff (during the interview) as an interview schedule must be followed.

In usability studies, unstructured interviews can be particularly useful in the early stages of mobile application development by identifying the pros and cons of graphical user interface design [122]. More generally, an unstructured interview can be used to elicit as many experiential statements as possible from a user after testing a product [123]. In addition, this method has been widely used to gather non-functional requirements, especially those that fall within the scope of usability [124].

A semi-structured interview format, which also relies on asking a series of open-ended questions, is said to elicit unbiased responses with the aim of uncovering usability issues by providing detailed qualitative data [125]. In practice, semi-structured interviews are used as a follow-up method, conducted either face-to-face or by email. It seems to be a widely accepted practice in the usability testing of mobile applications to combine both closed and open questions [126].

A structured interview is essentially the administration of a questionnaire, which ensures consistency and thoroughness [127]. The obvious advantage of a questionnaire is that it can be completed by the participant on paper or electronically, allowing relatively large samples of data to be collected with relatively little effort on the part of the experimenter, whereas disadvantages include inflexibility and the inability to pursue interesting lines of inquiry or to follow up on responses that may be unclear. However, the structured nature of such instruments allows them to be replicated across the research community, while testing their reliability and validity on different samples demonstrates their degree of generalizability.

It is worth noting that Fontana and Frey [128] provide a comprehensive overview of existing interview types, as well as useful guidelines for developing and conducting each one.

4.3. Usability Testing Process

In a general sense, a process is "a series of actions that produce something or that lead to a particular result" [129]. With this in mind, the process of mobile application usability testing consists of a sequence of three tasks:

1. Data collection.

2. Data analysis.
3. Usability assessment.

Each of these tasks can be considered as a separate part, thinking in terms of the workload required, including human resources, hardware equipment, tools, and methods. They are all discussed in more detail below.

4.3.1. Data Collection

The purpose of the data collection is to obtain all the primary data, necessary to (a) identify and describe the profile of the respondents, (b) analyze, reproduce, and evaluate the interaction, and (c) collect the feedback of the users, in terms of specific usability attributes and their metrics. By its very nature, a set of data is collected during a usability testing session, as shown in Figure 1.

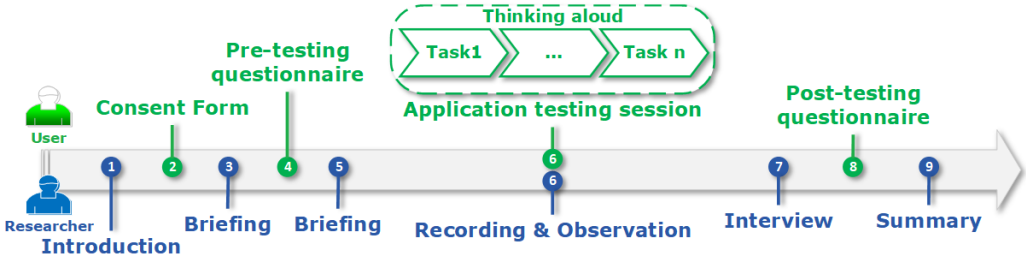

Figure 1. Data collection process.

As one can notice, an individual usability testing session involves two different actors: a researcher (marked in blue) and a user (marked in green). The course of the session proceeds in the following manner:

1. A session starts with an introduction. The researcher briefly presents the general assumptions, research objectives, research object and methods, session scenario (including tasks to be performed by the user), details of the thinking aloud protocol, components of the test environment, data collected and how the data will be used in the future, user rights, user confidentiality, and anonymity clauses.
2. The user is then asked to sign a consent form, a document that protects the rights of the participants, provides details of the above information, and ultimately builds trust between the researcher and the participants.
3. Next, the researcher briefly informs and instructs the participant about the content and purpose of the pre-testing questionnaire.
4. This instrument is used to collect demographic data, information about the participant's knowledge and skills in relation to the object of the study, and other relevant information. An example of a pre-testing questionnaire is available here [50].
5. During the second briefing, a researcher introduces a user with a pre-defined list of tasks to be performed by the user. All hardware equipment should be checked. Finally, if there are no obstacles, a user should be kindly asked to think aloud while performing each task.
6. The usability testing session is the core component of the data collection phase as voice and video data are collected using hardware equipment (document camera and microphone [130]). In addition, the researcher is also involved through monitoring and observation of the progress of the test session. From a practical perspective, written notes can often provide useful follow-up information.
7. Next, the participant is asked for additional feedback in the form of open questions on any unspoken or unobserved issues raised during the interaction with the application; this could take the form of an interview. If necessary, a short break afterwards is an option to consider.

8. Finally, the post-test questionnaire is submitted to a user. The aim is to collect primary data on the user's perceptions and experiences. An example of a post-test questionnaire can be found here [50].
9. In the summary, a researcher concludes the testing session and discusses any remaining organizational issues, if there are any.

At this point, it should also be emphasized that the pre-testing and post-testing questionnaires can be merged into a single questionnaire and thereby administered to a user after a testing session. In addition, the order in which the consent form is presented for the user to sign may be set differently.

4.3.2. Data Analysis

By its nature, data analysis is essentially an iterative process in which a researcher extracts the premises necessary to formulate conclusions. With this in mind, the analysis of collected data involves the following:

- video content analysis, which comprises annotation procedures including the user's actions and application responses, separated and marked on the timeline;
- identifying and documenting the application errors, defects, and malfunctions; and
- extracting all numerical values necessary to estimate particular attribute metrics.

It should be noted that it is common practice to use a video player application or other software tools to support this process. In addition, a variety of visualization techniques are usually used to facilitate the analysis and interpretation of the results obtained. For example, a timeline is a graphical method of displaying a list of a user's actions in chronological order.

4.3.3. Usability Assessment

Once all the usability measures have been measured and assessed, it is then possible to analyze, classify, and interpret the results. On the other hand, some may go through the audio–video recordings and study the user's task performance in detail. The result of this step is the report, which generally presents the results and conclusions, as well as a list of recommendations with the corresponding ratings of the participants. In particular, the report is divided into the following sections: (i) assessed usability attribute, along with their analysis and interpretation, (ii) bugs and errors, and (iii) recommendations and future research directions. Obviously, the structure and content of the report should correspond to both the goal and the formulated research questions. In this sense, the target audience may be testers, designers, and developers, respectively.

4.4. GAQM Framework

By definition, a research framework refers to the overall approach that combines conceptualizations and principles that serve as the basis for a phenomenon to be investigated [131]. More specifically, it is a systematic way of organizing and conceptualizing the research process, including the goal of the study, research questions and data collection methods that guide a research study. To design and develop our framework for mobile usability testing, the Goal–Question–Metric (GQM) approach was adopted and adapted [132] since it has been widely recognized due to its capacity for facilitating software quality [133].

By definition, GQM is based on goal orientation theory [134] and is designed in a top-down fashion. First, one must specify the rationale behind the measurement plan, which in turn must inform one of the goals. Second, questions are then derived to articulate the goal defined for an object. A goal should be expressed in terms of a measurable outcome. Next, each goal is broken down into at least one question, which should provide a definition for the measurement object with respect to a particular quality issue. Finally, one or more metrics are assigned to each question.

Based on this notion and the theoretical underpinnings discussed earlier, we introduce the Goal–Attribute–Question–Metric (GAQM) framework to structure, conceptualize, and operationalize the study of mobile usability testing. The GAQM defines a research process on four levels (see Figure 2):

1. Conceptual level (goal). The research goal is defined, including the usability dimension (observed or perceived) and the name of the mobile application (subject of the study); a goal could also refer to usability in general.
2. Contextual level (attribute). The mobile usability attributes are specified.
3. Operational level (question). At least one research question is formulated to operationalize each specific attribute.
4. Quantitative level (metric). At least one directly observable metric is assigned for an observed attribute, while two or more are assigned for a perceived attribute.

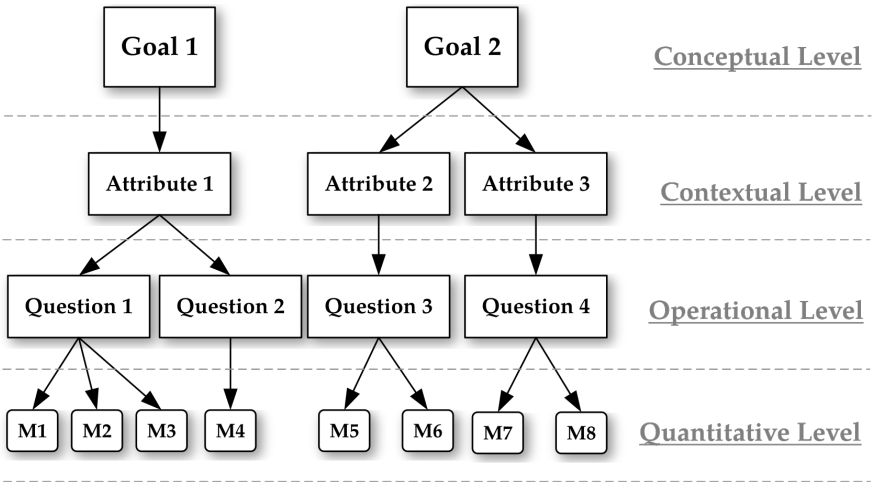

Figure 2. Illustration of the Goal–Attribute–Question–Metric (GAQM) framework.

We argue that the GCAM framework can be applied to any study of mobile usability testing. It is designed to clarify and emphasize both research assumptions, intentions, and measurements by providing a generic and structured approach. However, to support our view, the three use cases of the GAQM framework are discussed below.

Note that, the chosen usability testing environment (i.e., lab or field) should be able to address the mobile context (e.g., network connectivity) or the performance of specific application features (e.g., route updating). The context may be deliberately formulated, or it may be extracted from either the research objective or the specified tasks. In addition, any planned data collection techniques should also be explicitly stated.

5. Use Cases

The world has witnessed a significant shift in shopping behavior over the past decade. With the comfort of online shopping, consumers are turning more than ever to mobile applications to find and purchase products, and flowers are no exception.

Determining the layout of content is a sensitive task. Desktop devices share considerable larger screen space, whereas mobile devices are inherently limited. In fact, users are forced to view a small amount of content at a time before they have to scroll. A designer often struggles and wonders about the most efficient layout for the content presentation scheme (see Figure 3). Should a list view or a grid view be used? Undoubtedly, a decision can affect how quickly and easily users interact with the application.

List view presents content in a single-column list. It can be text-heavy, whereas an interface typically displays icons or thumbnails next to the text. App users rely on reading the information to make their choices. On the other hand, grid view displays content in two or more columns with images. The images dominate most of the space, and the text is truncated to avoid too much text wrapping. App users rely on the images to make their selections. Looking again at Figure 3, an obvious question arises: which of these two content schemas exhibits higher usability?

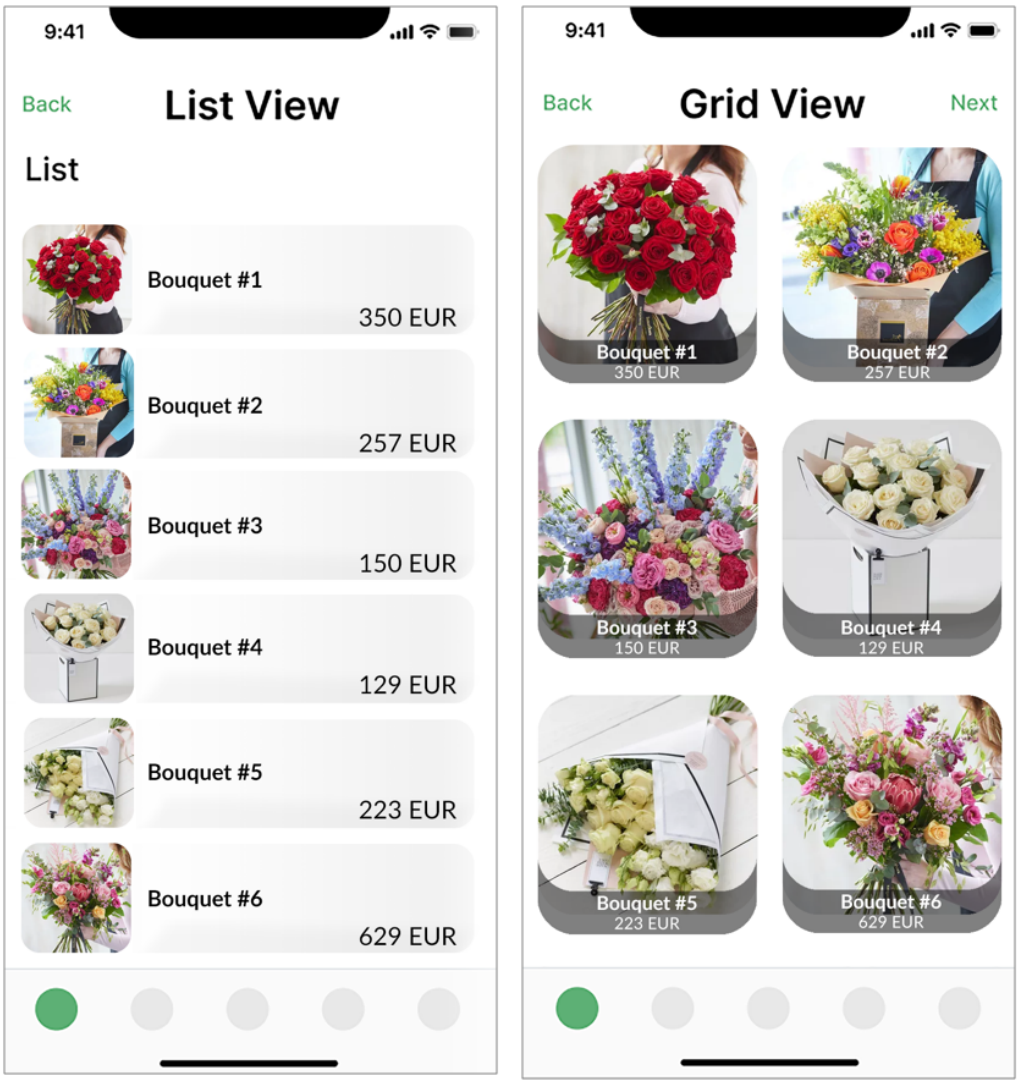

Figure 3. The high-fidelity prototypes of mobile applications for buying and sending flowers. On the left is a list view schema, while on the right is a grid view.

To demonstrate the value of the GAQM framework, we will look at three use cases, each in a different context. We will use the framework to structure the research process. In particular, to guide each of the hypothetical usability studies, we will formulate the research objective by directly decomposing it into the tangible usability attributes, together with the corresponding research questions, and assigning at least one metric to each of them.

5.1. Use Case #1

The first use case examines the choice between a list view and a grid view for a mobile e-commerce application for buying and sending flowers (hereafter referred to as Flying Flowers, or simply the application). This is illustrated in Figure 3, which shows two high-fidelity prototypes, applied separately in two independent implementations, denoted as version *A* and version *B*. It should also be assumed that, where applicable, analogous content schemas were used to design the rest of the user interface of each version. The research problem is to determine which version (*A* or *B*) has higher observable usability. To structure and guide the research process, the GAQM framework has been applied. The results are depicted by Figure 4.

In order to collect all the necessary data, a recorded testing session, separately for each version of the application, is carried out with individual users to collect video data, following the protocol shown in Figure 1. The extracted and estimated values of the metrics

are used to perform a Student's *t*-test to test the hypothesis of significant differences in the means of the observed effectiveness and efficiency between version *A* and version *B*. Based on this, an informed decision can be made as to which application version exhibits higher observable usability.

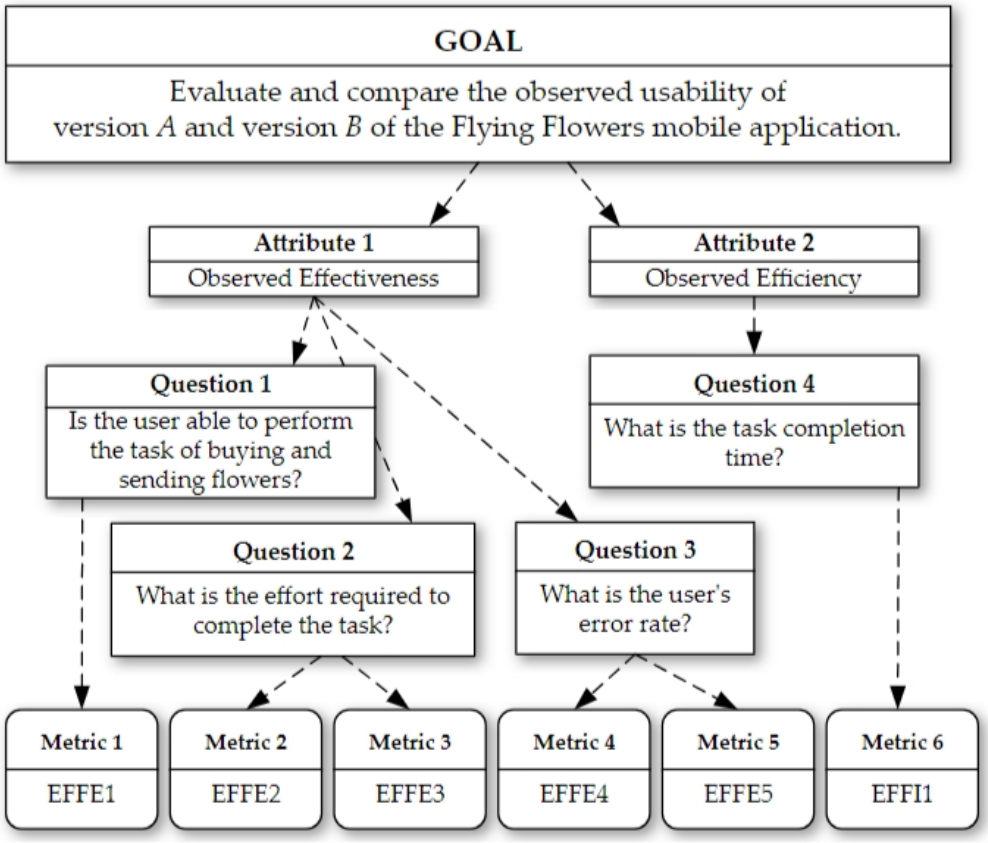

Figure 4. Use case of the GAQM framework for the observed usability study of version *A* and version *B* of the Flying Flowers app.

5.2. Use Case #2

In the second use case, we consider the similar research problem, as in the first use case. This time, however, usability is understood in terms of the perceived usability, which means that the current study aims to determine whether version *A* or version *B* of the Flying Flowers mobile application demonstrates higher perceived usability. More specifically, and thinking in terms of ISO 9241-11, perceived usability is understood in terms of three perceived attributes, namely effectiveness, efficiency, and satisfaction.

In a similar vein, in order to structure and guide the research process, the GAQM framework has been adopted. The results are depicted by Figure 5.

To collect quantitative data, a participant fills out the post-testing questionnaire after testing each version of the application. As can be seen, in the current settings, such a questionnaire contains at least 12 items since a total of 12 metrics have been assigned to all three usability attributes. The calculated values of the metrics are used to perform a Student's *t*-test to test the hypothesis of significant differences in the means of perceived effectiveness and efficiency between version *A* and version *B*. Based on this, an evidence-based decision can be made as to which version of the application has higher perceived usability.

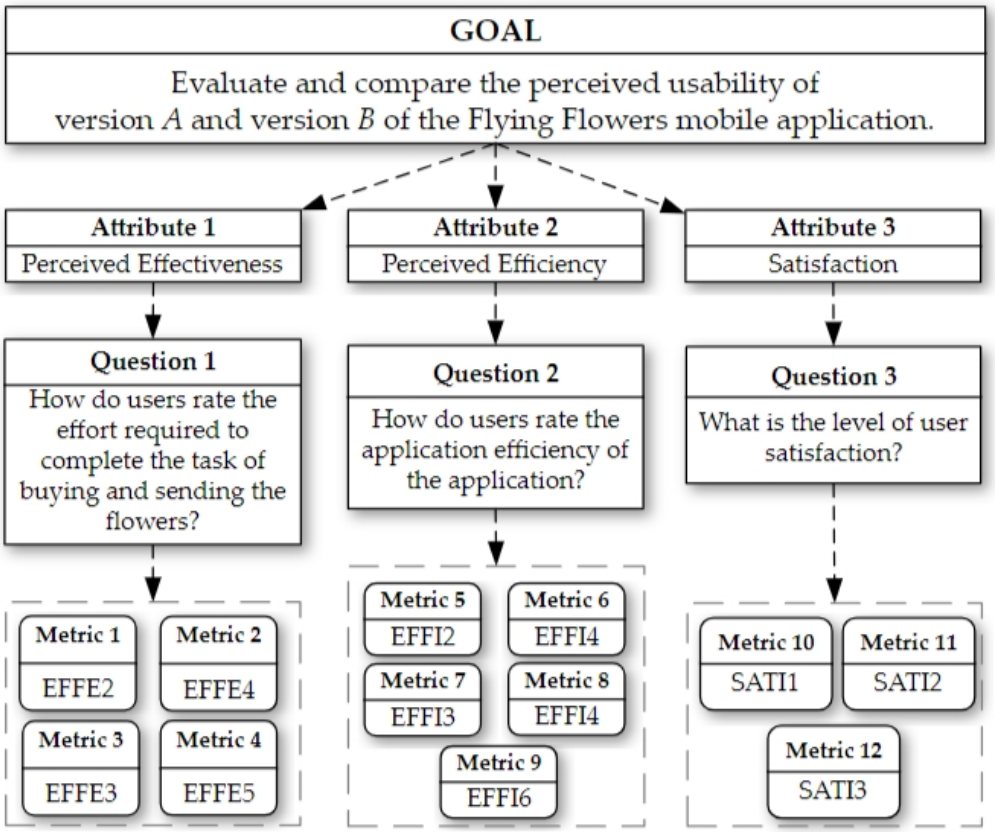

Figure 5. Use case of the GAQM framework for the perceived usability study of version *A* and version *B* of the Flying Flowers app.

5.3. Use Case #3

In the third use case, the research problem is similar to the other two discussed above. However, this time, usability is understood in general terms, which means that there is no need to distinguish its attributes. Therefore, at the conceptual level, only the goal of the study is formulated. At the remaining two levels, operational and quantitative, the research questions and metrics are defined in a similar way.

Furthermore, no usability testing session is designed and organized. Therefore, participants will be asked to install and use a version of *A* or a version of *B* on their own smartphones, alternately, at any time and at their own convenience. In such an approach, the 10-item System Usability Scale (SUS) with an adjective rating scale is used to measure and evaluate the usability of each version. Note that the SUS is claimed to be the most widely used measure of perceived usability [135].

In short, the current study aims to evaluate and compare version *A* and version *B* of the Flying Flowers mobile application. To organize, structure, and guide the research process, the GAQM framework was adopted and adapted to address both the study objective and settings. The results are depicted by Figure 6.

To collect all the necessary data, the link to the online survey is sent to the pool of users who will answer questions based on their experience. Alternatively, a paper-based questionnaire can be used, but the data collected must be transferred to a digital form. However, one should also consider verifying the experience level of each user by asking questions such as length of use, frequency of use over the past year, and overall satisfaction. Such simple measures allow a researcher to maintain the homogeneity of the sample, which increases the reliability of the study.

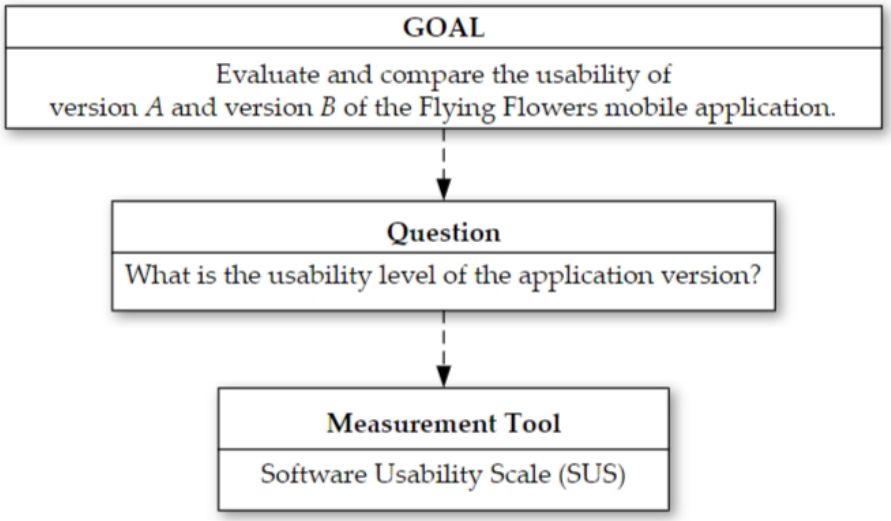

Figure 6. Use case of the GAQM framework for the usability study of version *A* and version *B* of the Flying Flowers app.

6. Discussion

Laboratory research is still more common than field research because the effort involved in organizing, preparing, testing, and analyzing participant samples is considerably less. In addition, a number of well-known usability measurement tools, such as the Software Usability Scale (SUS) or Software Usability Measurement Inventory (SUMI), can be easily adapted for English-speaking users, as well as translated into other languages and used to evaluate mobile usability [136,137]. This type of research, compared to the presented methodology, requires less time and money, but it provides only a general view of users regarding the pre-assigned aspects of the quality of mobile applications.

The advantages derived from mobile usability field studies are either authentic in nature, related to real conditions and settings, or reliable in nature, related to unmoderated design. According to Burghardt and Wirth [138], the need for participants to multitask during field testing significantly increases the error rate associated with task completion and thus the likelihood of identifying specific usability problems. On top of that, some user behavior can only be observed in the field [139], which captures the complexity and richness of the real world in which the interaction between a user and a mobile application is located [140]. In a broader sense, the value of field settings has also been recognized and confirmed by other studies related to the usability of mobile applications [141,142].

In the field of information systems research, the usefulness of questionnaires has long been recognized [143,144], and they have been widely used to study mobile usability. The use of a questionnaire as a research method is flexible and cost- and time-effective [145]. It allows for large-scale data collection distributed across a geographically diverse group of participants, as well as the standardization of data without the need for physical presence. While questionnaires are a popular and widely used research method [146], they have certain limitations. By design, questionnaires are often structured and may not allow participants to elaborate on their answers. In mobile usability testing, this can limit the depth of information gathered, making it difficult to gain a thorough understanding of complex issues [147].

Participant observation, in which a study is conducted through the researcher's direct participation in the usability testing, where the questions and discussions arise from the participant's involvement [148], is a qualitative method of considerable interest to the human–computer interaction community [149]. When considering mobile usability testing, participant observation allows researchers to gather in-depth and detailed information with the ability to explore different contexts by actively participating in the settings [150].

This allows a researcher to gain insights that may be difficult to capture through other research methods. Participant observation is feasible for both field and laboratory studies but appears to be more difficult to implement for the former approach.

In usability testing, thinking aloud is often used to assess how users interact with software, helping researchers to understand user concerns, preferences, and areas for design improvement [151]. Since participants express their thoughts as they occur, thinking aloud provides real-time data, allowing researchers to capture the immediate cognitive processes involved in performing a task [152], which appears to be particularly useful in A/B (split) testing. According to Nielsen [153], "thinking aloud may be the single most valuable usability engineering method". Currently, the method is widely used in usability testing of mobile applications [154].

As interviews allow information to be put into context, users can provide details about specific problems that have arisen. On the other hand, it is possible to elicit requirements that are not specified elsewhere [155]. If participants provide ambiguous or unclear answers, interviewers can ask for clarification in real time. However, participants may vary significantly in their ability to articulate thoughts and experiences. Some individuals may provide detailed and reflective responses, while others may struggle to express themselves, resulting in variability in the quality of the data [156]. While interviews have proven to be useful for mobile usability evaluation, uncovering various issues and revealing different user expectations [157], they can be time-consuming both in terms of preparation and actual data collection [158].

It would be at least naive to clearly state the best data collection technique for mobile usability testing. Since each has its own strengths and weaknesses, the most appropriate approach depends on the specified goals and attributes, as well as the testing context. In addition, factors such as the target users, the type of mobile application, the available resources, and the desired depth of insights play a vital role in determining the most appropriate data collection technique. Therefore, a preliminary evaluation of each method in relation to the study characteristics seems essential to make an informed decision.

On the other hand, the proposed process of data collection aims to structure and integrate different techniques, eliciting from a user both ad hoc (spur-of-the-moment) and ab illud (precise and conscious) information. In addition, video recording allows for retrospective analysis, which involves reviewing and reflecting on the usability testing sessions. However, this organized and guided approach requires considerable resources. Moreover, if prepared and executed correctly, designers and developers can make informed decisions to improve the overall user experience. It is especially important for newly developed mobile applications, which may be subject to final testing before market launch.

In terms of theoretical implications, our study has contributed to empirical studies by introducing the GAQM framework. This top-down approach can guide a researcher in formulating and communicating the design of a mobile usability testing study. By emphasizing the importance of aligning research goals with specific usability dimensions and considering the context in which the application is used, the framework provides a nuanced perspective on the multifaceted nature of mobile usability. More specifically, at this point and with the existing evidence in the previous literature, the novel aspects of the GAQM framework involve embodying of the two usability dimensions as foundations to conceptualize the design and implementation of a usability study of mobile applications.

In summary, the presented GAQM framework makes an important contribution to the development of theoretical perspectives and methodological approaches in the field of mobile application usability research. In our opinion, the GAQM is a good fit for this problem because it was relatively easy to map and decompose the research process into goals, attributes, questions, and metrics. In addition, the use cases presented can be easily adopted and adapted to other real-world settings, thus guiding other researchers in designing and conducting mobile usability studies.

Nevertheless, this study suffers from the limitations common to all qualitative research. First, the subjective nature of qualitative analysis introduces the potential obstacle of

individual bias as the researcher's prior knowledge and current understanding may influence the interpretation process. This threat was mitigated by rigorous reflexivity and transparency through the use of multiple sources of information established to inform a research community in an effective and reliable manner.

7. Conclusions

This paper explores the intricacies of usability testing for mobile applications as this area of research is burgeoning, largely due to the unique challenges posed by mobile devices, such as their unique characteristics, limited bandwidth, unreliable wireless networks, and the ever-changing context influenced by environmental factors. In particular, two categories of mobile usability are introduced, followed by the conceptualization of its three related attributes, namely effectiveness, efficiency, and satisfaction, borrowed from the ISO 9241-11 standard.

Our methodological framework begins with a discussion focused on an overview of research settings for usability testing of mobile applications, intended for laboratory and field studies, respectively. At the end, a short summary compares the two types. In this case, a first avenue for future research has emerged, which concerns the optimal site conditions. The research efforts undertaken could yield interesting and valuable findings, revealing new frontiers in mobile usability testing.

Afterwards, four different data collection techniques, including questionnaire, participant observation, thinking aloud, and interview, are described and analyzed. Thus, we introduce the reader to the key areas from the perspective of reliability and validity in any scientific research. More specifically, we discuss the specific settings, types, and favorable circumstances associated with each of these techniques. We believe that there is a need for more empirical research in this area to add more factual evidence to the current state of theory.

Next, the process of mobile usability testing is outlined. It consists of a sequence of three tasks: data collection, data analysis, and usability evaluation. From the point of view of organizing and conducting the research, the second task deserves special attention. Due to its unlimited scope and natural flexibility, this data collection scheme can be applied in any experimental setup, adapted to the needs and requirements imposed by the study objectives and the inherent context. Although there is still little research in this area, it is expected that more will be carried out in order to confirm the theoretical frameworks proposed in the current study.

Our methodological framework concludes with the introduction of a pragmatic approach designed to conceptualize and operationalize any study oriented towards mobile usability testing. The proposed GAQM framework is hierarchical in nature and operates on four entities that include goals, attributes, questions, and metrics. Similarly, we expect that the introduced framework will be appreciated by other researchers who, by its adoption, would confirm its applicability and external validity.

Funding: In June 2020, a new laboratory named "MobileUX" was established at the Faculty of Electronics, Telecommunications and Computer Science at the Gdansk University of Technology (GUT). It should be noted that the MobileUX laboratory was financed from the internal resources of the Faculty of Electronics, Telecommunications and Informatics.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: I would like to thank the former Dean, Jerzy Wtorek, for his financial support and commitment.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACM	Association for Computing Machinery
MDPI	Multidisciplinary Digital Publishing Institute
GAQM	Goal–Attribute–Question–Metric
GPS	Global Positioning System
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
UI	User Interface

References

1. Statista. Forecast Number of Mobile Users Worldwide from 2020 to 2025. 2023 . Available online: https://www.statista.com/statistics/218984/number-of-global-mobile-users-since-2010/ (accessed on 28 July 2023).
2. Statista. Time Spent with Nonvoice Activities on Mobile Phones Every Day in the United States from 2019 to 2024. 2023. Available online: https://www.statista.com/statistics/1045353/mobile-device-daily-usage-time-in-the-us/ (accessed on 28 July 2023).
3. Elite Content Marketer. Average Screen Time Statistics for 2023. 2023. Available online: https://elitecontentmarketer.com/screen-time-statistics/ (accessed on 28 July 2023).
4. Statista. Revenue from Smartphone Sales in the United States from 2013 to 2027. 2023. Available online: https://www.statista.com/statistics/619821/smartphone-sales-revenue-in-the-us/ (accessed on 28 July 2023).
5. Admiral Media. Why 99,5 percent of Consumer Apps Fail (And How To Keep Yours Alive). 2023. Available online: https://admiral.media/why-consumer-apps-fail-and-how-to-keep-yours-alive/ (accessed on 16 December 2023).
6. Goyal, A. Top Reasons Why Mobile Apps Fail to Make a Mark in the Market. 2019. Available online: https://www.businessofapps.com/insights/top-reasons-why-mobile-apps-fail-to-make-a-mark-in-the-market/ (accessed on 2 February 2024).
7. Swaid, S.I.; Suid, T.Z. Usability heuristics for M-commerce apps. In Advances in Usability, User Experience and Assistive Technology, Proceedings of the AHFE 2018 International Conferences on Usability & User Experience and Human Factors and Assistive Technology, Orlando, FL, USA, 21–25 July 2018; Springer: Berlin/Heidelberg, Germany, 2019; pp. 79–88.
8. Tode, C. More than Half of Consumers Dissatisfied with Mobile Retail Experiences. 2017. Available online: https://www.retaildive.com/ex/mobilecommercedaily/more-than-half-of-shoppers-are-dissatisfied-with-mobile-retail-experiences-adobe (accessed on 2 February 2024).
9. Hedegaard, S.; Simonsen, J.G. Extracting usability and user experience information from online user reviews. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 2089–2098.
10. Xu, G.; Gutiérrez, J.A. An exploratory study of killer applications and critical success factors in m-commerce. *J. Electron. Commer. Organ. (JECO)* **2006**, *4*, 63–79. [CrossRef]
11. Baharuddin, R.; Singh, D.; Razali, R. Usability dimensions for mobile applications-a review. *Res. J. Appl. Sci. Eng. Technol.* **2013**, *5*, 2225–2231. [CrossRef]
12. Desak, G.F.P.; Gintoro. List of most usability evaluation in mobile application: A systematic literature review. In Proceedings of the 2020 International Conference on Information Management and Technology (ICIMTech), Bandung, Indonesia, 13–14 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 283–287.
13. Au, F.T.; Baker, S.; Warren, I.; Dobbie, G. Automated usability testing framework. In Proceedings of the Ninth Conference on Australasian User Interface, Darlinghurst, NSW, Australia, 1 January 2008; Volume 76, pp. 55–64.
14. Joshua, S.R.; Abbas, W.; Lee, J.H.; Kim, S.K. Trust Components: An Analysis in The Development of Type 2 Diabetic Mellitus Mobile Application. *Appl. Sci.* **2023**, *13*, 1251. [CrossRef]
15. Hohmann, L. Usability: Happier users mean greater profits. *Inf. Syst. Manag.* **2003**, *20*, 66–76. [CrossRef]
16. Ahmad, W.F.W.; Sulaiman, S.; Johari, F.S. Usability Management System (USEMATE): A web-based automated system for managing usability testing systematically. In Proceedings of the 2010 International Conference on User Science and Engineering (i-USEr), Shah Alam, Malaysia, 13–15 December 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 110–115.
17. Cliquet, G.; Gonzalez, C.; Huré, E.; Picot-Coupey, K. From Mobile Phone to Smartphone: What's New About M-Shopping? In *Ideas in Marketing: Finding the New and Polishing the Old, Proceedings of the 2013 Academy of Marketing Science (AMS) Annual Conference, Monterey Bay, CA, USA, 15–18 May 2013*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 199–202.
18. Coursaris, C.K.; Kim, D.J. A meta-analytical review of empirical mobile usability studies. *J. Usability Stud.* **2011**, *6*, 117–171.
19. Thales. The Evolution of the Smartphone. 2023. Available online: https://www.thalesgroup.com/en/worldwide-digital-identity-and-security/mobile/magazine/evolution-smartphone (accessed on 17 December 2023).
20. Nacheva, R. Standardization issues of mobile usability. *Int. J. Interact. Mob. Technol.* **2020**, *14*, 149–157. [CrossRef]
21. Whittemore, R.; Knafl, K. The integrative review: Updated methodology. *J. Adv. Nurs.* **2005**, *52*, 546–553. [CrossRef]
22. Zhang, D.; Adipat, B. Challenges, methodologies, and issues in the usability testing of mobile applications. *Int. J.-Hum.-Comput. Interact.* **2005**, *18*, 293–308. [CrossRef]

23. Ji, Y.G.; Park, J.H.; Lee, C.; Yun, M.H. A usability checklist for the usability evaluation of mobile phone user interface. *Int. J.-Hum.-Comput. Interact.* **2006**, *20*, 207–231. [CrossRef]
24. Heo, J.; Ham, D.H.; Park, S.; Song, C.; Yoon, W.C. A framework for evaluating the usability of mobile phones based on multi-level, hierarchical model of usability factors. *Interact. Comput.* **2009**, *21*, 263–275. [CrossRef]
25. Hussain, A.; Kutar, M. Usability metric framework for mobile phone application. In Proceedings of the PG Net'09: 10th Annual Conference on the Convergence of Telecommunications, Networking and Broadcasting, Liverpool, UK, 22–23 June 2009.
26. Jeong, J.; Kim, N.; In, H.P. Detecting usability problems in mobile applications on the basis of dissimilarity in user behavior. *Int. J.-Hum.-Comput. Stud.* **2020**, *139*, 102364. [CrossRef]
27. Weichbroth, P. Usability of mobile applications: A systematic literature study. *IEEE Access* **2020**, *8*, 55563–55577. [CrossRef]
28. *ISO 9241-11:2018*; Ergonomics of huMan-System Interaction—Part 11: Usability: Definitions and Concepts. International Organization for Standardization, Geneva, Switzerland, 2018.
29. Owoc, M.; Weichbroth, P.; Żuralski, K. Towards better understanding of context-aware knowledge transformation. In Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 3–6 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1123–1126.
30. Yuan, H.; Jin, T.; Ye, X. Establishment and Application of Crowd-Sensing-Based System for Bridge Structural Crack Detection. *Appl. Sci.* **2023**, *13*, 8281. [CrossRef]
31. Rafique, U.; Khan, S.; Ahmed, M.M.; Kiani, S.H.; Abbas, S.M.; Saeed, S.I.; Alibakhshikenari, M.; Dalarsson, M. Uni-planar MIMO antenna for sub-6 GHz 5G mobile phone applications. *Appl. Sci.* **2022**, *12*, 3746. [CrossRef]
32. Nakhimovsky, Y.; Miller, A.T.; Dimopoulos, T.; Siliski, M. Behind the scenes of google maps navigation: Enabling actionable user feedback at scale. In Proceedings of the CHI'10 Extended Abstracts on Human Factors in Computing Systems, Atlanta, CA, USA, 10–15 April 2010; pp. 3763–3768.
33. Musumba, G.W.; Nyongesa, H.O. Context awareness in mobile computing: A review. *Int. J. Mach. Learn. Appl.* **2013**, *2*, 5. [CrossRef]
34. Luo, C.; Goncalves, J.; Velloso, E.; Kostakos, V. A survey of context simulation for testing mobile context-aware applications. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–39. [CrossRef]
35. Encyclopedia.com. Attribute. 2024. Available online: https://www.encyclopedia.com/science-and-technology/computers-and-electrical-engineering/computers-and-computing/attribute (accessed on 20 January 2024).
36. Zaibon, S.B. User testing on game usability, mobility, playability, and learning content of mobile game-based learning. *J. Teknol.* **2015**, *77*, 131–139. [CrossRef]
37. Gilbert, A.L.; Sangwan, S.; Ian, H.H.M. Beyond usability: The OoBE dynamics of mobile data services markets. *Pers. Ubiquitous Comput.* **2005**, *9*, 198–208. [CrossRef]
38. Silvennoinen, J.; Vogel, M.; Kujala, S. Experiencing visual usability and aesthetics in two mobile application contexts. *J. Usability Stud.* **2014**, *10*, 46–62.
39. Widyanti, A.; Ainizzamani, S.A.Q. Usability evaluation of online transportation'user interface. In Proceedings of the 2017 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, 23–24 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 82–86.
40. Harrison, R.; Flood, D.; Duce, D. Usability of mobile applications: Literature review and rationale for a new usability model. *J. Interact. Sci.* **2013**, *1*, 1–16. [CrossRef]
41. John, B.E.; Marks, S.J. Tracking the effectiveness of usability evaluation methods. *Behav. Inf. Technol.* **1997**, *16*, 188–202. [CrossRef]
42. Jeng, J. Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *Libri* **2005**, *55*, 96–121. [CrossRef]
43. Kabir, M.A.; Salem, O.A.; Rehman, M.U. Discovering knowledge from mobile application users for usability improvement: A fuzzy association rule mining approach. In Proceedings of the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 24–26 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 126–129.
44. Baumeister, R.F. *Encyclopedia of Social Psychology*; Sage: London, UK, 2007; Volume 1.
45. Choi, W.; Stvilia, B. Web credibility assessment: Conceptualization, operationalization, variability, and models. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 2399–2414. [CrossRef]
46. Carneiro, A. Maturity and metrics in health organizations information systems. In *Handbook of Research on ICTs and Management Systems for Improving Efficiency in Healthcare and Social Care*; IGI Global: Hershey, PA, USA, 2013; pp. 937–952.
47. United States Agency for International Development. Glossary of Evaluation Terms. 2009. Available online: https://pdf.usaid.gov/pdf_docs/PNADO820.pdf (accessed on 20 January 2024).
48. Bollen, K.A.; Diamantopoulos, A. In defense of causal-formative indicators: A minority report. *Psychol. Methods* **2017**, *22*, 581. [CrossRef] [PubMed]
49. Edwards, J.R.; Bagozzi, R.P. On the nature and direction of relationships between constructs and measures. *Psychol. Methods* **2000**, *5*, 155. [CrossRef] [PubMed]
50. Weichbroth, P. An empirical study on the impact of gender on mobile applications usability. *IEEE Access* **2022**, *10*, 119419–119436. [CrossRef]
51. Cambridge Dictionary. Satisfaction. 2024. Available online: https://dictionary.cambridge.org/dictionary/english/satisfaction (accessed on 20 January 2024).

52. Liébana-Cabanillas, F.; Molinillo, S.; Ruiz-Montañez, M. To use or not to use, that is the question: Analysis of the determining factors for using NFC mobile payment systems in public transportation. *Technol. Forecast. Soc. Chang.* **2019**, *139*, 266–276. [CrossRef]
53. Hsiao, C.H.; Chang, J.J.; Tang, K.Y. Exploring the influential factors in continuance usage of mobile social Apps: Satisfaction, habit, and customer value perspectives. *Telemat. Inform.* **2016**, *33*, 342–355. [CrossRef]
54. Wan, L.; Xie, S.; Shu, A. Toward an understanding of university students' continued intention to use MOOCs: When UTAUT model meets TTF model. *Sage Open* **2020**, *10*, 2158244020941858. [CrossRef]
55. Lodhi, A. Usability heuristics as an assessment parameter: For performing usability testing. In Proceedings of the 2010 2nd International Conference on Software Technology and Engineering, San Juan, PR, USA, 3–5 October 2010; IEEE: Piscataway, NJ, USA, 2010; Volume 2, pp. V2-256–V2-259.
56. Chisman, J.; Diller, K.; Walbridge, S. Usability testing: A case study. *Coll. Res. Libr.* **1999**, *60*, 552–569. [CrossRef]
57. Wichansky, A.M. Usability testing in 2000 and beyond. *Ergonomics* **2000**, *43*, 998–1006. [CrossRef]
58. Lewis, J.R. Usability testing. In *Handbook of Human Factors and Ergonomics*; John Wiley & Sons: Hoboken, NJ, USA, 2012; pp. 1267–1312.
59. Riihiaho, S. Usability testing. In *The Wiley Handbook of Human Computer Interaction*; John Wiley & Sons: Hoboken, NJ, USA, 2018; Volume 1, pp. 255–275.
60. Mason, P.; Plimmer, B. A critical comparison of usability testing methodologies. *NACCQ* **2005**, 255–258. Available online: https://www.cs.auckland.ac.nz/~beryl/publications/NACCQ%202005%20Critical%20Comparison%20of%20Usability%20Testing%20Methodologies.pdf (accessed on 17 December 2023).
61. Gaffney, G. Information & Design Designing for Humans. 1999. Available online: https://infodesign.com.au/assets/UsabilityTesting.pdf (accessed on 28 July 2023).
62. Dumas, J.S.; Redish, J. *A Practical Guide to Usability Testing*; Intellect Books: Bristol, UK, 1999.
63. Sauer, J.; Sonderegger, A.; Heyden, K.; Biller, J.; Klotz, J.; Uebelbacher, A. Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing. *Appl. Ergon.* **2019**, *74*, 85–96. [CrossRef] [PubMed]
64. Betiol, A.H.; de Abreu Cybis, W. Usability testing of mobile devices: A comparison of three approaches. In Proceedings of the IFIP Conference on Human-Computer Interaction, Orlando, FL, USA, 9–14 July 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 470–481.
65. Gawlik-Kobylińska, M.; Kabashkin, I.; Misnevs, B.; Maciejewski, P. Education Mobility as a Service: A Study of the Features of a Novel Mobility Platform. *Appl. Sci.* **2023**, *13*, 5245. [CrossRef]
66. Dehlinger, J.; Dixon, J. Mobile application software engineering: Challenges and research directions. In Proceedings of the Workshop on Mobile Software Engineering, Honolulu, HI, USA, 22–24 May 2011; Volume 2, pp. 29–32.
67. Schusteritsch, R.; Wei, C.Y.; LaRosa, M. Towards the perfect infrastructure for usability testing on mobile devices. In Proceedings of the CHI'07 extended abstracts on Human Factors in Computing Systems, San Jose, CA, USA, 28 April–3 May 2007; pp. 1839–1844.
68. Genaidy, A.M.; Karwowski, W. The emerging field of health engineering. *Theor. Issues Ergon. Sci.* **2006**, *7*, 169–179. [CrossRef]
69. Keith, M.; Shao, B.; Steinbart, P.J. The usability of passphrases for authentication: An empirical field study. *Int. J.-Hum.-Comput. Stud.* **2007**, *65*, 17–28. [CrossRef]
70. Aziz, H.A. Comparison between field research and controlled laboratory research. *Arch. Clin. Biomed. Res.* **2017**, *1*, 101–104. [CrossRef]
71. Harbach, M.; Von Zezschwitz, E.; Fichtner, A.; De Luca, A.; Smith, M. It's a hard lock life: A field study of smartphone (Un)Locking behavior and risk perception. In Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS 2014), Menlo Park, CA, USA, 9–1 July 2014; pp. 213–230.
72. Pousttchi, K.; Thurnher, B. Understanding effects and determinants of mobile support tools: A usability-centered field study on IT service technicians. In Proceedings of the 2006 International Conference on Mobile Business, Copenhagen, Denmark, 26–27 June 2006; IEEE: Piscataway, NJ, USA, 2006; p. 10.
73. Van Elzakker, C.P.; Delikostidis, I.; van Oosterom, P.J. Field-based usability evaluation methodology for mobile geo-applications. *Cartogr. J.* **2008**, *45*, 139–149. [CrossRef]
74. Kjeldskov, J.; Graham, C.; Pedell, S.; Vetere, F.; Howard, S.; Balbo, S.; Davies, J. Evaluating the usability of a mobile guide: The influence of location, participants and resources. *Behav. Inf. Technol.* **2005**, *24*, 51–65. [CrossRef]
75. Pensabe-Rodriguez, A.; Lopez-Dominguez, E.; Hernandez-Velazquez, Y.; Dominguez-Isidro, S.; De-la Calleja, J. Context-aware mobile learning system: Usability assessment based on a field study. *Telemat. Inform.* **2020**, *48*, 101346. [CrossRef]
76. Rowley, D.E. Usability testing in the field: Bringing the laboratory to the user. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 24–28 April 1994; pp. 252–257.
77. Kallio, T.; Kaikkonen, A.; Cankar, M.; Kallio, T.; Kankainen, A. Usability testing of mobile applications: A comparison between laboratory and field testing. *J. Usability Stud.* **2005**, *1*, 23–28.
78. Nielsen, C.M.; Overgaard, M.; Pedersen, M.B.; Stage, J.; Stenild, S. It's worth the hassle! the added value of evaluating the usability of mobile systems in the field. In Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles, Oslo, Norway, 14–18 October 2006; pp. 272–280.

79. Duh, H.B.L.; Tan, G.C.; Chen, V.H.h. Usability evaluation for mobile device: A comparison of laboratory and field tests. In Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services, Amsterdam, The Netherlands, 2–5 September 2006; pp. 181–186.
80. Nayebi, F.; Desharnais, J.M.; Abran, A. The state of the art of mobile application usability evaluation. In Proceedings of the 2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Montreal, QC, Canada, 29 April–2 May 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–4.
81. Kaya, A.; Ozturk, R.; Altin Gumussoy, C. Usability measurement of mobile applications with system usability scale (SUS). In Proceedings of the Industrial Engineering in the Big Data Era: Selected Papers from the Global Joint Conference on Industrial Engineering and Its Application Areas, GJCIE 2018, Nevsehir, Turkey, 21–22 June 2018; Springer: Berlin/Heidelberg, Germany, 2019; pp. 389–400.
82. Suzuki, S.; Bellotti, V.; Yee, N.; John, B.E.; Nakao, Y.; Asahi, T.; Fukuzumi, S. Variation in importance of time-on-task with familiarity with mobile phone models. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver BC Canada, 7–12 May 2011; pp. 2551–2554.
83. Kugler, S.; Czwick, C.; Anderl, R. Development of a valuation method for IoT-platforms. In Proceedings of the Product Lifecycle Management in the Digital Twin Era: 16th IFIP WG 5.1 International Conference, PLM 2019, Moscow, Russia, 8–12 July 2019; Revised Selected Papers 16; Springer: Berlin/Heidelberg, Germany, 2019; pp. 293–301.
84. Partala, T.; Saari, T. Understanding the most influential user experiences in successful and unsuccessful technology adoptions. *Comput. Hum. Behav.* **2015**, *53*, 381–395. [CrossRef]
85. Sonderegger, A.; Schmutz, S.; Sauer, J. The influence of age in usability testing. *Appl. Ergon.* **2016**, *52*, 291–300. [CrossRef] [PubMed]
86. Alturki, R.; Gay, V. Usability testing of fitness mobile application: Methodology and quantitative results. *Comput. Sci. Inf. Technol* **2017**, *7*, 97–114.
87. Ahmad, N.A.N.; Hussaini, M. A Usability Testing of a Higher Education Mobile Application Among Postgraduate and Undergraduate Students. *Int. J. Interact. Mob. Technol.* **2021**, *15*. [CrossRef]
88. Cambridge Dictionary. Meaning of Questionnaire in English. 2023. Available online: https://dictionary.cambridge.org/dictionary/english-polish/questionnaire (accessed on 6 August 2023).
89. Cambridge Dictionary. Meaning of Survey in English. 2023. Available online: https://dictionary.cambridge.org/dictionary/english-polish/survey (accessed on 6 August 2023).
90. Nielsen Norman Group. Open-Ended vs. Closed-Ended Questions in User Research. 2016. Available online: https://www.nngroup.com/articles/open-ended-questions/ (accessed on 6 August 2023).
91. Sauro, J.; Kindlund, E. A method to standardize usability metrics into a single score. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, OR, USA, 2–7 April 2005; pp. 401–409.
92. Ponto, J. Understanding and evaluating survey research. *J. Adv. Pract. Oncol.* **2015**, *6*, 168.
93. Preissle, J. Participant Observation. In *The Corsini Encyclopedia of Psychology*; University of Michigan: Ann Arbor, MI, USA, 2010; pp. 1–2.
94. Ahmed, A.A.; Muhammad, R.A. Participant observation of a farmers-herders community in Anguwar Jaba Keffi Nigeria. *Int. J. Sci. Res. Publ.* **2021**, *11*, 84–87. [CrossRef]
95. Musante, K.; De Walt, B.R. *Participant Observation as a Data Collection Method*; Rowman Altamira: Lanham, MD, USA, 2010.
96. Dorazio, P.; Stovall, J. Research in context: Ethnographic usability. *J. Tech. Writ. Commun.* **1997**, *27*, 57–67. [CrossRef]
97. Kawulich, B.B. Participant observation as a data collection method. *Forum Qual. Sozialforschung Forum Qual. Soc. Res.* **2005**, *6*, 43. [CrossRef]
98. Jorgensen, D.L. *Participant Observation: A Methodology for Human Studies*; Sage: London, UK, 1989; Volume 15.
99. Tomlin, W.C.; Tomlin, W.C. UX and Usability Testing Data. In *UX Optimization: Combining Behavioral UX and Usability Testing Data to Optimize Websites*; Apress: New York, NY, USA, 2018; pp. 97–127.
100. Mohamad, U.H.; Abdul Hakim, I.N.; Ali-Akbari, M. Usability of a gamified antibiotic resistance awareness mobile application: A qualitative evaluation. *IET Netw.* **2022**. [CrossRef]
101. Khayyatkhoshnevis, P.; Tillberg, S.; Latimer, E.; Aubry, T.; Fisher, A.; Mago, V. Comparison of Moderated and Unmoderated Remote Usability Sessions for Web-Based Simulation Software: A Randomized Controlled Trial. In Proceedings of the International Conference on Human-Computer Interaction, Virtual, 26 June–1 July 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 232–251.
102. Kamińska, D.; Zwoliński, G.; Laska-Leśniewicz, A. Usability Testing of Virtual Reality Applications—The Pilot Study. *Sensors* **2022**, *22*, 1342. [CrossRef] [PubMed]
103. Fisher, E.A.; Wright, V.H. Improving online course design through usability testing. *J. Online Learn. Teach.* **2010**, *6*, 228–245.
104. Svanæs, D.; Alsos, O.A.; Dahl, Y. Usability testing of mobile ICT for clinical settings: Methodological and practical challenges. *Int. J. Med. Inform.* **2010**, *79*, e24–e34. [CrossRef] [PubMed]
105. Güss, C.D. What is going through your mind? Thinking aloud as a method in cross-cultural psychology. *Front. Psychol.* **2018**, *9*, 1292. [CrossRef] [PubMed]
106. Trickett, S.; Trafton, J.G. A primer on verbal protocol analysis. In *The PSI Handbook of Virtual Environments for Training and Rducation*; Bloomsbury Academic: New York, NY, USA, 2009; Volume 1, pp. 332–346.

107. Ericsson, K.A.; Simon, H.A. Protocol Analysis (Revised Edition); *Overview of Methodology of Protocol Analysis*; MIT Press: Cambridge, MA, USA, 1993.
108. Bernardini, S. Think-aloud protocols in translation research: Achievements, limits, future prospects. *Target. Int. J. Transl. Stud.* **2001**, *13*, 241–263. [CrossRef]
109. Jensen, J.J. Evaluating in a healthcare setting: A comparison between concurrent and retrospective verbalisation. In Proceedings of the International Conference on Human-Computer Interaction, Beijing, China, 22–27 July 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 508–516.
110. Fan, M.; Lin, J.; Chung, C.; Truong, K.N. Concurrent think-aloud verbalizations and usability problems. *ACM Trans.-Comput.-Hum. Interact. (TOCHI)* **2019**, *26*, 1–35. [CrossRef]
111. Hertzum, M.; Borlund, P.; Kristoffersen, K.B. What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *Int. J.-Hum.-Comput. Interact.* **2015**, *31*, 557–570. [CrossRef]
112. Hertzum, M.; Holmegaard, K.D. Thinking aloud in the presence of interruptions and time constraints. *Int. J.-Hum.-Comput. Interact.* **2013**, *29*, 351–364. [CrossRef]
113. Siegel, D.; Dray, S. A Usability Test Is Not an Interview. *ACM Interact.* **2016**, *23*, 82–84.
114. Boren, T.; Ramey, J. Thinking aloud: Reconciling theory and practice. *IEEE Trans. Prof. Commun.* **2000**, *43*, 261–278. [CrossRef]
115. Hertzum, M.; Hansen, K.D.; Andersen, H.H. Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behav. Inf. Technol.* **2009**, *28*, 165–181. [CrossRef]
116. Nasruddin, Z.A.; Markom, A.; Abdul Aziz, M. Evaluating construction defect mobile app using think aloud. In Proceedings of the User Science and Engineering: 5th International Conference, i-USEr 2018, Puchong, Malaysia, 28–30 August 2018; Proceedings 5; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
117. Conway, J.M.; Peneno, G.M. Comparing structured interview question types: Construct validity and applicant reactions. *J. Bus. Psychol.* **1999**, *13*, 485–506. [CrossRef]
118. Chauhan, R.S. Unstructured interviews: Are they really all that bad? *Hum. Resour. Dev. Int.* **2022**, *25*, 474–487. [CrossRef]
119. Galletta, A. *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication*; NYU Press: New York, NY, USA, 2013; Volume 18.
120. Kallio, H.; Pietilä, A.M.; Johnson, M.; Kangasniemi, M. Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. *J. Adv. Nurs.* **2016**, *72*, 2954–2965. [CrossRef]
121. Wilson, J.L.; Hareendran, A.; Hendry, A.; Potter, J.; Bone, I.; Muir, K.W. Reliability of the modified Rankin Scale across multiple raters: Benefits of a structured interview. *Stroke* **2005**, *36*, 777–781. [CrossRef]
122. Ormeño, Y.I.; Panach, J.I.; Pastor, O. An Empirical Experiment of a Usability Requirements Elicitation Method to Design GUIs based on Interviews. *Inf. Softw. Technol.* **2023**, *164*, 107324. [CrossRef]
123. Lindgaard, G.; Dudek, C. User satisfaction, aesthetics and usability: Beyond reductionism. In *Usability: Gaining a Competitive Edge*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 231–246.
124. Liang, L.; Tang, Y.; Tang, N. Determinants of groupware usability for community care collaboration. In Proceedings of the Frontiers of WWW Research and Development-APWeb 2006: 8th Asia-Pacific Web Conference, Harbin, China, 16–18 January 2006; Proceedings 8; Springer: Berlin/Heidelberg, Germany, 2006; pp. 511–520.
125. Walji, M.F.; Kalenderian, E.; Piotrowski, M.; Tran, D.; Kookal, K.K.; Tokede, O.; White, J.M.; Vaderhobli, R.; Ramoni, R.; Stark, P.C.; et al. Are three methods better than one? A comparative assessment of usability evaluation methods in an EHR. *Int. J. Med. Inform.* **2014**, *83*, 361–367. [CrossRef] [PubMed]
126. Jiang, T.; Luo, G.; Wang, Z.; Yu, W. Research into influencing factors in user experiences of university mobile libraries based on mobile learning mode. *Libr. Hi Tech* 2022, *ahead-of-print*. [CrossRef]
127. Stanton, N.A.; Hedge, A.; Brookhuis, K.; Salas, E.; Hendrick, H.W. *Handbook of Human Factors and Ergonomics Methods*; CRC Press: Boca Raton, FL, USA, 2004.
128. Fontana, A.; Frey, J.H. The interview. In *The Sage Handbook of Qualitative Research*; Sage: London, UK, 2005; Volume 3, pp. 695–727.
129. Britannica Dictionary. Britannica Dictionary Definition of PROCESS. 2023. Available online: https://www.britannica.com/dictionary/process (accessed on 1 August 2023).
130. Budiu, R. Usability Testing for Mobile Is Easy. 2014. Available online: https://www.nngroup.com/articles/mobile-usability-testing/ (accessed on 9 January 2024).
131. Salkind, N.J. *Encyclopedia of Research Design*; Sage: London, UK, 2010; Volume 1.
132. Caldiera, V.R.B.G.; Rombach, H.D. The goal question metric approach. In *Encyclopedia of Software Engineering*; University of Michigan: Ann Arbor, MI, USA, 1994; pp. 528–532.
133. Nick, M.; Althoff, K.D.; Tautz, C. Facilitating the practical evaluation of organizational memories using the goal-question-metric technique. In Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management, Dagstuhl Castle, Germany, 26–29 May 1999.
134. Kaplan, A.; Maehr, M.L. The contributions and prospects of goal orientation theory. *Educ. Psychol. Rev.* **2007**, *19*, 141–184. [CrossRef]
135. Ramli, R.Z.; Wan Husin, W.Z.; Elaklouk, A.M.; Sahari@ Ashaari, N. Augmented reality: A systematic review between usability and learning experience. *Interact. Learn. Environ.* **2023**, 1–17. [CrossRef]

136. Akmal Muhamat, N.; Hasan, R.; Saddki, N.; Mohd Arshad, M.R.; Ahmad, M. Development and usability testing of mobile application on diet and oral health. *PLoS ONE* **2021**, *16*, e0257035. [CrossRef]
137. Zaini, H.; Ishak, N.H.; Johari, N.F.M.; Rashid, N.A.M.; Hamzah, H. Evaluation of a Child Immunization Schedule Application using the Software Usability Measurement Inventory (SUMI) Model. In Proceedings of the 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET), Shah Alam, Malaysia, 6 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 281–285.
138. Burghardt, D.; Wirth, K. Comparison of evaluation methods for field-based usability studies of mobile map applications. In Proceedings of the International Cartographic Conference, Paris, France, 3–8 July 2011.
139. Taniar, D. *Mobile Computing: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*; IGI Global: Hershey, PA, USA, 2008; Volume 1.
140. López-Gil, J.M.; Urretavizcaya, M.; Losada, B.; Fernández-Castro, I. Integrating field studies in agile development to evaluate usability on context dependant mobile applications. In Proceedings of the XV International Conference on Human Computer Interaction, Puerto de la Cruz, Spain, 10–21 September 2014; pp. 1–8.
141. Von Zezschwitz, E.; Dunphy, P.; De Luca, A. Patterns in the wild: A field study of the usability of pattern and pin-based authentication on mobile devices. In Proceedings of the 15th international Conference on Human-Computer Interaction with Mobile Devices and Services, Munich, Germany, 27–30 September 2013; pp. 261–270.
142. Knip, F.; Bikar, C.; Pfister, B.; Opitz, B.; Sztyler, T.; Jess, M.; Scherp, A. A field study on the usability of a nearby search app for finding and exploring places and events. In Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia, Melbourne, VIC, Australia, 25–28 November 2014; pp. 123–132.
143. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
144. Davis, W.S.; Yen, D.C. *The Information System Consultant's Handbook: Systems Analysis and Design*; CRC Press: Boca Raton, FL, USA, 2019.
145. Fife-Schaw, C. Questionnaire design. In *Research Methods in Psychology*; SAGE: An Caisteal Nuadh, UK, 1995; pp. 174–193.
146. Maramba, I.; Chatterjee, A.; Newman, C. Methods of usability testing in the development of eHealth applications: A scoping review. *Int. J. Med. Inform.* **2019**, *126*, 95–104. [CrossRef]
147. Lim, K.C.; Selamat, A.; Alias, R.A.; Krejcar, O.; Fujita, H. Usability measures in mobile-based augmented reality learning applications: A systematic review. *Appl. Sci.* **2019**, *9*, 2718. [CrossRef]
148. Panchea, A.M.; Todam Nguepnang, N.; Kairy, D.; Ferland, F. Usability Evaluation of the SmartWheeler through Qualitative and Quantitative Studies. *Sensors* **2022**, *22*, 5627. [CrossRef]
149. Romero-Ternero, M.; García-Robles, R.; Cagigas-Muñiz, D.; Rivera-Romero, O.; Romero-Ternero, M. Participant Observation to Apply an Empirical Method of Codesign with Children. *Adv.-Hum.-Comput. Interact.* **2022**, *2022*, 1–5. [CrossRef]
150. Álvarez Robles, T.d.J.; Sánchez Orea, A.; Álvarez Rodríguez, F.J. UbicaME, mobile geolocation system for blind people: User experience (UX) evaluation. *Univers. Access Inf. Soc.* **2023**, *22*, 1163–1173. [CrossRef]
151. McDonald, S.; Edwards, H.M.; Zhao, T. Exploring think-alouds in usability testing: An international survey. *IEEE Trans. Prof. Commun.* **2012**, *55*, 2–19. [CrossRef]
152. Van Waes, L. Thinking aloud as a method for testing the usability of websites: The influence of task variation on the evaluation of hypertext. *IEEE Trans. Prof. Commun.* **2000**, *43*, 279–291. [CrossRef]
153. Nielsen, J. *Usability Engineering*; Morgan Kaufmann: Burlington, MA, USA, 1994.
154. Borys, M.; Milosz, M. Mobile application usability testing in quasi-real conditions. In Proceedings of the 2015 8th International Conference on Human System Interaction (HSI), Warsaw, Poland, 25–27 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 381–387.
155. Garmer, K.; Ylvén, J.; Karlsson, I.M. User participation in requirements elicitation comparing focus group interviews and usability tests for eliciting usability requirements for medical equipment: A case study. *Int. J. Ind. Ergon.* **2004**, *33*, 85–98. [CrossRef]
156. Roulston, K. Considering quality in qualitative interviewing. *Qual. Res.* **2010**, *10*, 199–228. [CrossRef]
157. Hussain, A.; Mkpojiogu, E.O.; Ishak, N.; Mokhtar, N.; Ani, Z.C. An Interview Report on Users' Perception about the Usability Performance of a Mobile E-Government Application. *Int. J. Interact. Mob. Technol.* **2019**, *13*, 169–178. [CrossRef]
158. Sarkar, U.; Gourley, G.I.; Lyles, C.R.; Tieu, L.; Clarity, C.; Newmark, L.; Singh, K.; Bates, D.W. Usability of commercially available mobile applications for diverse patients. *J. Gen. Intern. Med.* **2016**, *31*, 1417–1426. [CrossRef]