Multimodal Technologies and Interaction

MDPI

Article

Cross-Platform Usability Model Evaluation

Khalid Majrashi [1,*], Margaret Hamilton [2], Alexandra L. Uitdenbogerd [2] and Shiroq Al-Megren [3]

[1] Department of Information Technology, Institute of Public Administration, Riyadh 11141, Saudi Arabia
[2] School of Science (Computer Science), Royal Melbourne Institute of Technology (RMIT) University, Melbourne 3000, Australia; margaret.hamilton@rmit.edu.au (M.H.); sandra.uitdenbogerd@rmit.edu.au (A.L.U.)
[3] Mechanical Engineering Department, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA; shiroq@mit.edu
* Correspondence: MajrashiK@ipa.edu.sa

Received: 25 August 2020; Accepted: 13 November 2020; Published: 20 November 2020

Abstract: It is becoming common for several devices to be utilised together to access and manipulate shared information spaces and migrate tasks between devices. Despite the increased worldwide use of cross-platform services, there is limited research into how cross-platform service usability can be assessed. This paper presents a novel cross-platform usability model. The model employs the think-aloud protocol, observations, and questionnaires to reveal cross-platform usability problems. Two Likert scales were developed for measuring overall user satisfaction of cross-platform usability and user satisfaction with the seamlessness of the transition between one device and another. The paper further employs a series of objective measures for the proposed model. The viability and performance of the model were examined in the context of evaluating three cross-platform services across three devices. The results demonstrate that the model is a valuable method for assessing and quantifying cross-platform usability. The findings were thoroughly analysed and discussed, and subsequently used to refine the model. The model was also evaluated by eight user experience experts and seven out of the eight agreed that it is useful.

Keywords: cross-platform; multi-device; user study; usability model; cross-device

1. Introduction

The world is imbued with technologies that have created and continue to create rapid change that manifests itself in all aspects of our lives [1]. The past two decades have hailed computer revolutions that have changed various aspects of its role in people's lives, from undertaking routine activities (e.g., paying bills or shopping) to creating wholly new experiences (e.g., virtual and augmented realities) [2]. One of the most dramatic developments in computing power is the global proliferation of mobile devices (smart phones and tablets) as convenient auxiliaries to the conventional desktop and personal computers. This collection of tools (personal computer and mobile devices) are further supported by the advent of the cloud-computing paradigm where users are able to access services from different devices and enabling interactions across these devices [3–5]. This vision for ubiquitous computing is not new [6] as the evolution from single-device to multi-device computing has long been predicted [7]. For instance, a user may draft an email on their phone and proceed to their personal computer to include an attachment; essentially achieving their goal *horizontally* across devices or platforms.

As people possess more information devices, it has become common for several devices to be used together as doorways into a shared information space [4,5,7–9]. Interaction across a blended cross-device ecology is commonly referred to as cross-device interaction, where users are enabled to manipulate shared content via multiple separate input and output devices within a perceived

interaction space [10]. Cross-device interaction can occur in various modes, such as moving sequentially from one device to another at different times (sequential interaction), and interacting simultaneously with more than one device at the same time (simultaneous interaction) [4,7]. Of these two modes of interaction, sequential interaction is the most common as it reflects current user interaction with multiple devices [4]; according to a Think with Google report, 90% of users ($n = 1611$) use multiple devices sequentially to accomplish a task, where 98% of users move between their devices at different times in a single day.

In Human–Computer Interaction (HCI) and Software Engineering communities, usability testing is a method of usability assessment that is used to assess an individual product by testing it on representative users in representative environments [11–13]. Usability testing utilises measurable factors, such as efficiency and effectiveness, to assess real user performance [14]. Usability has traditionally been an important quality attribute of interactive systems, where usable interfaces were found to improve human productivity and performance [12,15,16]. The growth of cross-platform interaction and services has resulted in a new emerging theme for usability referred to as 'cross-platform usability' (or 'inter-usability' or 'horizontal usability') (see [17–23]).

Although the traditional usability testing method is valuable for assessing product usability, it only involves quantifying and assessing the usability of a single user interface. This could be seen as the main limitation of service usability engineering. That is, the current testing method does not address how to quantify and assess inter-usability across a combination of user interfaces that involves a user transferring from one interface to another to achieve interrelated goals. The evaluation of each user interface independently might not correspond with the evaluation of a combination of user interfaces where users migrate tasks across platforms [17,24]. This paper aims to address this gap with the development of an assessment model for evaluating the cross-platform usability of a combination of user interfaces. The model's validity and performance was explored via a cross-platform usability test of three test services. The findings of the test were analysed and utilised to refine the proposed model.

The remainder of this paper is organised as follows. Section 2 defines concepts related to cross-platform services and configuration. Next, Section 3 reviews the literature for prior work in the area of cross-platform modelling and interaction. Section 4 presented the proposed cross-platform usability assessment model. The following section, Section 5, describes the user study conducted to assess the viability of the proposed model. Then, Section 6 presents the results of the user study. Section 7 discusses the results and refines the proposed assessment model. Section 8 presents experts' evaluations of the assessment model. Finally, Section 9 concludes the paper and briefly discusses future work.

2. Cross-Platform Services and Configuration

Interactive cross-platform systems are known in the literature by several terms, including "multiple user interfaces" (MUIs), which is defined as the view of the same information and services accessed by users from different computing platforms (hardware and software) [25–28]. The terms "multiple platform user interface" [29,30], "distributed user interface" (DUI) [31–33], "multi-channelling" and "cross media" [34] have also been used to describe interactive cross-platform systems. For this paper, we adopt the term "cross-platform service" to refer to a set of user interfaces for a single service on two or more computational platforms. This term is specifically used to highlight the transitions from one platform to another to complete tasks.

The configuration of a cross-platform service describes the way by which the multiple devices are organised [35]. Denis and Karsenty [17] outlined three degrees of device redundancy with respect to data and function availability across devices: redundant, complementary, and exclusive devices. Redundant devices allow access to the same data and functions either independently or responsively. With complementary devices, the cross-device user interfaces share a zone of data and function, but one or more of the devices provide access to data or functions that are inaccessible on other devices. An exclusive degree of cross-device user interface ensures unique access to different data and functions.

In this paper, the proposed model is intended to evaluate cross-platform usability of services that are configured at a complementary level of redundancy.

3. Related Work

In the past few decades, there has been a rapid and drastic change in the way we interact with computers as they become more powerful and easily accessible. This has brought the research in the area of cross-platform user interfaces and interactions into the limelight of Computing and Usability studies [5]. Research addressing the opportunities and challenges of cross-platform interaction is abundant and variable in contribution to HCI research [36]. Brudy et al. [5] provides an overview of cross-device research trends and terminology to unify common understanding for future research. A number of researchers have contributed to the body of knowledge concerning the design and development of cross-device systems (e.g., [21,36–40]). Dong et al. [21] explored the barriers of designing and developing multi-device experiences. Through a series of interviews with designers and developers of cross-device systems, the authors identified three challenges pertaining to designing interactions, complexity of user interface standards, and lack of evaluation tools and techniques. Other work, such as O'Leary et al. [37] and Sanchez-Adame et al. [38], provide designers toolkits and guidelines for cross-device user interfaces for context shifts and consistency, respectively. These studies reflect the relevance of the work presented in the paper, and for the rest of this section we explore the evaluation methods previously utilised in the area of cross-device evaluations.

Denis and Karsenty [17] studied the inter-usability of cross-device systems, whereby users migrate their tasks from one device to another. Their work found that service continuity could influence inter-device transitions. They argue that service continuity could have two different dimensions: knowledge continuity and task continuity. Their findings suggest that design principles such as inter-device consistency can be applied to user interface designs to support service continuity dimensions. However, although their work established an initial conceptual framework with inter-usability principles to support continuity, there was not a methodological approach for measuring the continuity factor in either a subjective or an objective way.

Seamless transition is an important user experience element in multi-device interaction mode. Dearman and Pierce [41] studied the techniques that people use to access multiple devices to produce a better user experience (UX) when working across devices. They argue that one of the main challenges is supporting seamless device changes. Concerning seamless transfer between devices, a few studies have attempted to address the issue of migrating tasks or applications across devices to reduce the impacts of interruptions when transitioning from one device to another and support continuity (e.g., [42,43]). However, these studies generally focused more on the technological aspects of the problem [27]. Dearman and Pierce [41] suggest that there are opportunities to improve the cross-platform UX by focusing on users rather than on applications and devices. Hence, our research aim is to provide a user-based testing methodology that leads to insights into user experiences and needs for task continuity and seamless transition.

Antila and Lui [24] investigated challenges that designers might encounter when designing inter-usable systems in the emerging field of interactive systems. Semi-structured interviews were conducted with 17 professionals working on interaction design in different domains. Challenges were identified and grouped by design phases: early, design, development, and evaluation phases. One of the identifiable challenges in the evaluation phase was the difficulty of evaluating the whole interconnected system. That is, the usability evaluation of separate components might not necessarily correspond to the evaluation of inter-usability. The work also reported the need for evaluation methods and metrics to support inter-usability, taking into account two factors: the composition of functionalities and the continuity of interaction. Therefore, this paper aims to address this need by providing a model for assessing inter-usability taking into consideration the cross-platform interaction factors.

Many of the related studies investigating cross-platform usability and UX, such as Wäljas et al. [35] and Denis and Karsenty [17], did not conduct user testing. Instead, they relied on data-collection

methods such as interviews and diaries that may not have been the most reliable sources for research in usability and UX. The interview method is generally used as a supplement to other usability research methods such as user testing. During interviews, people may not be able to remember the details of how they used a specific user interface; further, participants tend to make up stories to rationalise their behaviour and make it sound more logical than it may, in actual fact, have been. In addition, many users have no idea how to categorise their uses of technology according to a description. This means that what users say and what they do could be different. Therefore, a user-based testing method is still required for a reliable assessment of cross-platform service usability.

4. Cross-Platform Usability Assessment Model

A novel cross-platform usability assessment model is developed to enable the identification and quantification of cross-platform usability issues (see Figure 1). The key model elements are described in the following subsections: horizontal tasks (see Section 4.1); a collection of data collection techniques (see Section 4.2); cross-platform usability and seamless transition scales (see Sections 4.3 and 4.4); cross-platform usability metrics (see Section 4.5).

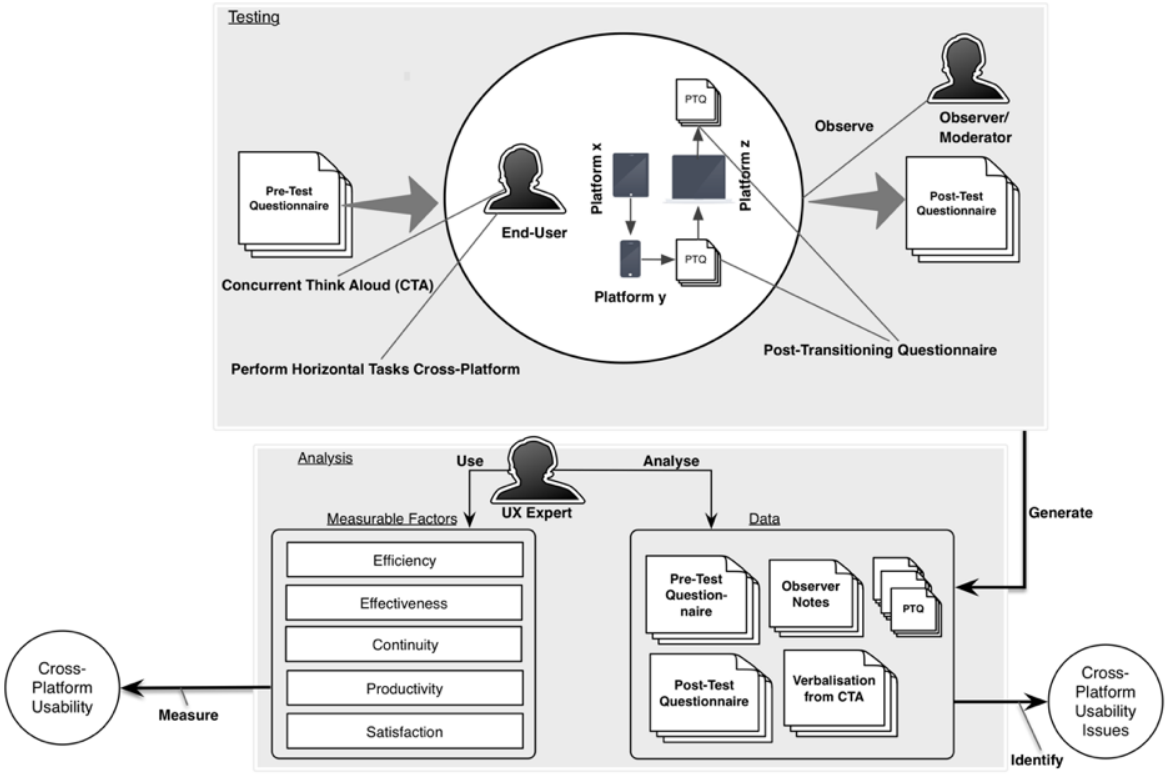

Figure 1. The proposed cross-platform usability assessment model for three user interfaces.

4.1. Horizontal Tasks

In the context of cross-platform usability testing, a cross-platform task needs to be divided into subtasks, each of which must be conducted using a specific user interface. In this paper, each set of related subtasks is referred to as a horizontal task (HT). The division of HTs was made to reflect the real tasks that a user performs in daily life across platforms. A horizontal goal was created for each HT that users are required to achieve. The horizontal goal is similarly the goal of the last subtask to be conducted.

Figure 2 illustrates a conceptual view of a HT that is achieved using two user interfaces as a representation for multiple user interfaces. The user starts a HT with the first user interface (e.g., locating a news article). The user then switches to the second user interface and performs

the second part of the task (e.g., adding comments to the news article), which is the final goal of the HT (the horizontal goal). The number of related subtasks that represent a single HT depends on the number of user interfaces involved in the test.

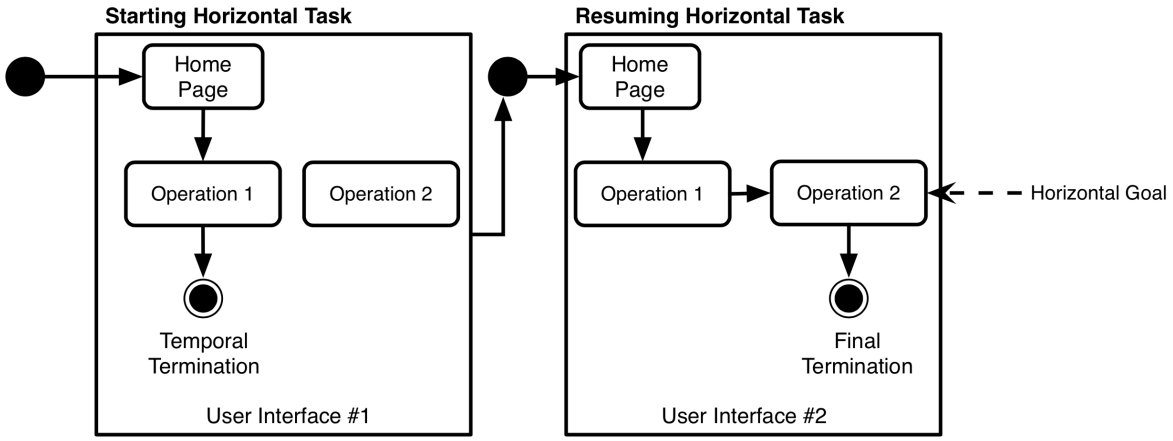

Figure 2. A conceptual view of a horizontal task.

4.2. Data Collection Techniques

It is common practice in usability testing to combine data gathering techniques to elicit usability issues from users [11]. In developing the proposed cross-platform usability assessment model, three techniques are adopted: concurrent think-aloud protocol, observations, and questionnaires. The proposed model utilises the benefits of each technique to produce a comprehensive assessment of cross-platform usability by identifying the largest number of cross-platform usability issues.

The think-aloud protocol is employed in usability research to collect a user's thoughts when interacting with systems. Users are asked to verbalise their thoughts while they attempt to accomplish tasks using a specific user interface. The think-aloud protocol is widely used among usability evaluators and is frequently used as the main source of data for usability testing, both in laboratory and field settings [44]. There are two main types of think-aloud protocols: concurrent (CTA) and retrospective (RTA). CTA expects users to verbalise their thoughts while carrying out tasks, while RTA collects users' thoughts after completing tasks via video recording of their task performance. Recent findings show that CTA is able to detect a higher number of usability problems and is more successful at facilitating successful usability tests [45].

Observations are utilised in usability studies to gather data about user context, tasks, and goals [46]. The techniques involve watching users while they interact with the interface. In usability tests, the main goal of an observer can be to note usability problems, record the time spent on a task, and discern tasks' successes or failures [47]. With observation, participants are observed directly by an evaluator who uses protocols to record observations. Alternatively, observations can be carried out indirectly via live video recordings or screen captures. Direct observation has the advantage of capturing the details of user interaction behaviour; however, user activity can be disrupted. In indirect observations, participants will work normally with fewer interruptions of their activities. However, fewer details of user interaction behaviour can be captured.

The proposed model utilises a pre-test questionnaire to collect demographic data, details of cross-platform expertise (e.g., periods of using devices and experiences with the tested service from each device), and user expectation of the distributed content. Post-test questionnaires in usability studies can provide diagnostic information about usability problems and measure user satisfaction immediately after the completion of a task, which can increase validity [48].

4.3. Cross-Platform Usability Scale

There are several standard product usability scales used by academic and industry researchers, including the System Usability Scale (SUS) [49], the Computer System Usability Questionnaire (CSUQ) [50], and the Questionnaire for User Interface Satisfaction (QUIS) [14]. The available system usability scales, while invaluable, only measure the user satisfaction of a single user interface. Consequently, a new Cross-Platform Usability Scale (CPUS) is developed and utilised in the proposed assessment model. CPUS draws inspiration from SUS and CSUQ, as well as from inter-usability design principles identified in prior work (e.g., inter-device consistency [17]).

The design of the CPUS followed the standard practice of the Likert scale [51] by using an equal number of negative and positive statements in random order. For positive questionnaire statements, a score of five consistently indicates a strongly positive attitude, while a score of one implies a strongly negative attitude. For negative questionnaire statements, a score of one indicates a strongly positive attitude, while five reveals a strongly negative attitude. The questionnaire consists of eight statements addressing a range of usability attributes pertaining to cross-platform user interaction: productivity, ease of use, expectation, degree of improvement, integration, learnability, consistency and frustration. The eight statements are:

1. I felt productive when using many platforms.
2. It was easy to use each user interface.
3. I found each system cross-platform designed in the way I expected it.
4. I felt that user interfaces cross-platform needed much improvement.
5. I found the various functions cross-platform were well integrated.
6. I needed to learn how to use each user interface separately.
7. I noticed inconsistencies between user interfaces cross-platform.
8. I was frustrated by the different designs of each user interface.

4.4. Seamless Transition Scale

The proposed assessment model also utilised a Seamless Transition Scale (STS) to measure user satisfaction with the seamlessness of transition between devices based on the attribute of 'continuity' pertaining to cross-platform usability [17]. Users are expected to rate the seamless transition immediately after undertaking two sequential tasks that require moving from one device to another. The STS comprises three statements that are concerned with measuring the seamless transition between user interfaces (continuity on a task) as perceived by the user:

- I am satisfied with the amount of time it took to resume the task I started from the (e.g., mobile device).
- I found I needed to remember information from the user interface on the (e.g., mobile device) to be able to continue with the task using the user interface on the (e.g., tablet).
- I felt I could seamlessly continue with my task after switching from the user interface on the (e.g., mobile device) to the user interface on the (e.g., tablet).

The STS was developed separately from the CPUS for two main reasons: (1) the assessment anticipates multiple transitions such that delaying the collection of data with the CPUS may confuse the users; and (2) collecting data after each transition supports measuring the transition of each transferred-to-user interface separately, or for each specific order of user interfaces.

4.5. Cross-Platform Usability Metrics

The cross-platform assessment model adopts three conventional measurable factors of usability: efficiency, effectiveness, and satisfaction [52]. Productivity and continuity are two additional factors used in the model. Productivity has previously been used as an attribute of usability in terms of user productivity when attempting a goal (e.g., [53]). In contrast to efficiency, productivity is concerned

with the amount of useful output obtained through user interaction [12]. Continuity is a new factor that is the most pertinent to cross-platform usability [17,54]. Table 1 lists the overall metrics of the proposed model and their definitions.

Table 1. Definition of cross-platform usability metrics.

Metric	Parameters	Settings
Efficiency	Execution time	The time spent executing a HT.
	Actions	The total number of steps required to complete a HT.
Effectiveness	Task completion	The successful completion of a HT.
	Errors	The total number of errors when progressing towards a HT goal.
Continuity	Task resuming success	The successful continuation of an interrupted task after switching devices.
	Resuming time	The time a user needs to resume an interrupted task after transition.
	Seamless transition	The score produced by the STS.
Productivity	Unproductive period	The time spent seeking help or recovery from errors when progressing towards a HT goal.
Satisfaction	Satisfaction	The score produced by the CPUS.

Figure 3 illustrates the cross-platform usability metrics utilised in the proposed cross-platform assessment model. Two metrics for measuring efficiency are used: cross-platform (horizontal) task execution time and action counts. The task completion rate and error count are employed to measure effectiveness. To measure continuity, the following metrics are used: task resuming success, the time taken to resume a task, and the score of seamless transition (generated though the STS). Unproductive periods are viewed as a single metric under the productivity factor. The scores generated via the CPUS are used to measure user satisfaction.

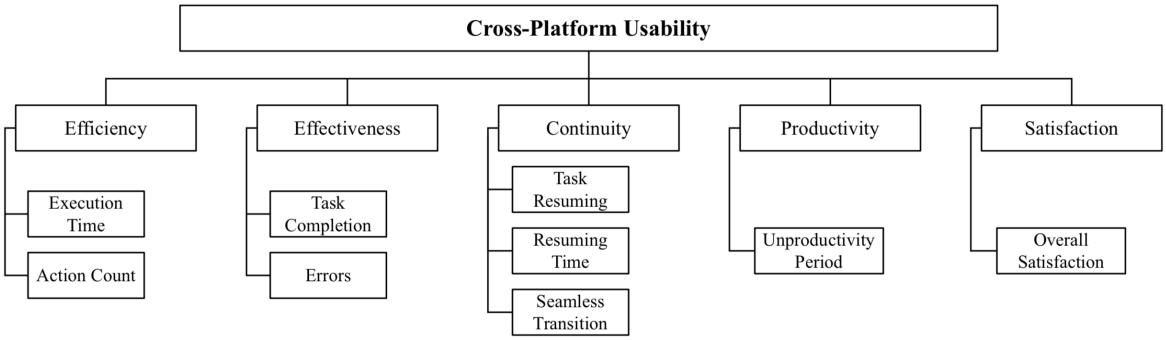

Figure 3. Cross-platform usability metrics.

5. User Study

The main purpose of this study is to address the following question: to what extent is the proposed cross-platform assessment model valuable for assessing cross-platform usability? The proposed model was developed to enable the identification of cross-platform usability issues and to quantify cross-platform usability. The main components of the model can be summarised as the use of HTs, the incorporation of a mix of data collection techniques, the consideration of cross-platform usability and seamless transition satisfaction scales, and the use of cross-platform usability objective metrics.

The user study will explore the following aspects of the proposed model:

- The extent to which the data-collection techniques proposed in our model help to reveal cross-platform usability issues.

- The extent to which the metrics can provide valuable information supporting cross-platform usability assessment.
- The correlation between metric groups under efficiency, effectiveness, and continuity attributed in the model.
- The correlation between statements on each Likert scale used in the model, the reliability of each scale, and the investigation of whether the statements on each scale as a whole could reflect only one dimension.

5.1. Test Objects and Tasks

The study used three cross-platform services from different domains of lifestyle, travel, and education: Real Estate (www.realestate.com.au), Trip Advisor (www.tripadvisor.com), and TED (www.ted.com). The services are composed of an array of features that correlated with activities often performed cross-platform, such as searching for information and planning a trip [4]. The services were configured at a complementary level of redundancy. The services have a range of different implementation types across platforms, that is, native mobile or tablet application, desktop website, mobile website, and responsive website. A native application is an application that is coded in a device-specific programming language. A desktop website is a web-based user interface that is optimised to be accessible from large screens (e.g., desktop and laptop screen). A mobile website is a copy of the desktop website, where the server delivers an optimised page that is smaller and easier to navigate on a smaller mobile screen. With a responsive website, the device automatically adjusts the site according to a device's screen size (large or small) and orientation (landscape or portrait).

The devices used in this study were a MacBook Pro-15 inch, an Apple iPad Air, an Apple iPhone 4, and a Samsung Galaxy S4. These devices belong to three main device categories (PC/laptop, tablet and smartphone) that people use in their daily lives [4]. These device categories are also used commonly in cross-platform sequential interactions [4], the interaction mode adopted in our experiments.

Real Estate (RE) lists properties for sale and rent in different areas of Australia. Users can search or browse properties by entering a specific area name or postcode or by selecting their current location. Users interacted with the RE service using a desktop website on the MacBook Pro, native application on the iPad, and mobile website on the Samsung Galaxy S4.

Trip Advisor (TA) offers advice from real travellers and a wide variety of travel choices. Users can search or browse different travel choices by entering a city or selecting their current location. Users interacted with the TA service using a desktop website on the MacBook pro, native application on the iPad, and native application on the Samsung Galaxy S4.

TED (TD) is a cross-platform service for spreading ideas in the form of short talks. Users can search or browse short talks and sort them according to different options, as well as filter search results. Users interacted with the TD service using a responsive website on the MacBook pro, native application on the iPad, and responsive website on the iPhone 4.

After determining the test services, a set of HTs were developed to assess the usability of the chosen objects by means of the proposed cross-platform assessment model. Six HTs were designed, two for each of the cross-platform services. We reiterate that the purpose of the study is to explore the viability of the proposed model and not to cover all the features of examined test objects. All HTs were designed to be carried out independently from each other. The tasks were piloted with one participant for each of the cross-platform services (i.e., three participants overall) prior to the commencement of data collection. This brings the total number of participants down to thirteen users (four participants for TA and TD, and five participants for RE). Table 2 lists the six HTs and subtasks adopted in this study.

5.2. Participants

What constitutes an optimal number of participants for usability testing has long been debated in the field [55–61]. Commonly, yet controversially, researchers have stated that four to nine participants is an adequate number to carry out an effective usability test. For that purpose and due to the

lack of consensus, it was decided that four users would be the minimal number considered for the study. For this study, sixteen participants were recruited and distributed amongst the three cross-device services.

An important consideration for usability participants was taken into account, that is, they are representative of the target user groups for the services being evaluated. This approach ensures valid feedback in order to contribute meaningful insight and explore the viability of the proposed model. Considering the services under evaluation, the study sample was selected from among university students. The age of the recruited participants was 18 to 60 years old at varying levels of study at the university. All participants had basic computer skills and used the internet on a daily basis for more than three years. The majority of participants had actively engaged in cross-device interaction for at least a year.

Table 2. Horizontal tasks for the test services.

Service	HT	Subtasks
RE	HT1	1. Find a property for rent in Reservoir (postcode: 3073) with two rooms, one toilet and two car spaces. 2. Find the inspection time for the property found in the previous interaction. 3. Find the contact information for the agent of the property.
	HT2	1. Find all sold properties in Preston (postcode: 3072). 2. Sort all sold properties in Preston from the lowest to the highest prices. 3. Sort all sold properties in Preston from the most to the least recently sold.
TA	HT1	1. Find all available three-star hotels in Sydney from 5 January 2015 to 13 January 2015. 2. Save a three-star hotel in Sydney. 3. Save another three-star hotel in Sydney.
	HT2	1. Find an Asian restaurant in Sydney with an average meal price of less than $100. 2. Show the restaurant found in the previous interaction on the map. 3. Write the following review about the restaurant: 'It is a good restaurant'.
TD	HT1	1. Find the talk, 'How not to be ignorant about the world'. 2. Share the talk, 'How not to be ignorant about the world', with a friend via email. 3. Share the talk, 'How not to be ignorant about the world', with your Facebook friends.
	HT2	1. Find the most viewed talks. 2. Save one of the most viewed talks. 3. Save another of the most viewed talks.

5.3. Procedure

The user study was conducted in the usability laboratory at the Royal Melbourne Institute of Technology (RMIT). Participants were cordially greeted upon arrival by the moderator (first author) and made to feel at ease. The participants were then asked to review an information sheet and sign an informed consent form. Participants were given no training in the selected cross-platform services or in the use of the devices involved in the study. However, they received some explanation about the think-aloud protocol, the terms used in the test session, and the main purpose of each cross-platform service. Participants were divided into three groups using matched-group design, through which the subjects are matched according to particular variables (e.g., age) and then allocated into groups. Each group performed tasks on a single cross-platform service. Each participant attempted two HTs. To achieve a HT, the participant was required to interact with three user interfaces for the same service accessed from three different devices: a laptop, a tablet, and a mobile phone. Test sessions were on average 85 minutes long. The sessions were video-recorded and the devices were also recorded to capture users' interactions with the devices.

The participants were introduced to the test service, and the moderator set up the screen capture software and video cameras. The participants were first asked to answer the pre-test questionnaire for the purpose of collecting demographic information. Participants then commenced performing each of the HTs for their assigned service. With each HT, participants performed subtask 1 and 2 on

the first two devices followed by the STS. Next, participants performed subtask 3 on the third device followed by the second STS. After concluding the first HT, the participants proceeded to the second HT following the same procedure. After all tasks were completed, the moderator ended the recording and directed the participant to fill in the CPUS to conclude the session.

To minimise order effects that could potentially bias the results of the study, a basic Latin square design was used to alternate the device order and tasks. This design resulted in a block of six trials with different orders. Changing the device order also allowed for the assessment of cross-platform usability in each specific order.

5.4. Usability Problem Extraction

The proposed cross-platform assessment model utilised a combination of techniques for data collection, which includes concurrent think aloud protocol, observation, and questionnaires (see Section 4). For the concurrent think aloud protocol, participants were encouraged to narrate their thoughts while interacting with the devices in the study. This approach is popularly adopted in usability studies as it allows a better understanding of participants' levels of engagement, their thoughts and feelings during interaction, and any questions that may arise. Since the protocol is applied simultaneously while the test is being carried out, it saves time and can help assist the participants to organise their thoughts while interacting with the set tasks [45,62].

Usability issues were analysed by counting the number of problems that those three methods were able to identify [63–66]. The process of identifying usability problems in this study involved reviewing each participant's testing video, observation notes, and questionnaires. The analysis order effect was reduced by randomly selecting data files. Statements were extracted from user verbalisation, post-transitioning, post-test questionnaires and observation notes. Each usability problem discovered was assigned a number that indicates the participants and assigned test service. Usability problems were maintained in a report that also contained information on context.

6. Results

This section presents the findings of the user study pertaining to data collection techniques, factors affecting cross-platform usability measures, and cross-platform usability metrics grouped under the proposed model's factors: efficiency, effectiveness, continuity, productivity, and satisfaction.

6.1. Data Collection Techniques

The proposed model utilised a combination of techniques for data collection (see Section 4). Usability problems were analysed by counting the number of problems, which is a common analysis technique [65,66]. In total, 540 cross-platform usability issues were identified. For all three services, the average number of usability issues was the highest for the think-aloud protocol (33.0, 26.5 and 21.75, respectively). For RE this is followed next by STS, CPUS, and observations (8.6, 3.6, and 2.4). In the case of the TA service, the think-aloud protocol was followed by observations, STS, and CPUS (9.3, 7.0 and 2.8). For TD, next follows CPUS, observations, and then STS (4.0, 2.8, and 1.3).

The think-aloud protocol revealed more cross-platform usability issues. Nevertheless, the other data collection techniques contributed to the process of identifying usability issues. It was found that some of the issues reported through questionnaires and observations were actually different from those verbalised using the think-aloud protocol. That is, 50%, 70%, and 69% of the issues identified through observation in the RE, TA, and TD services, respectively, were unique. With the post-test questionnaire, 45%, 67%, and 40% of the usability issues revealed in the RE, TA, and TD services, respectively, were different to those from other methods. A further 50%, 54%, and 34% of the issues discovered by the STS for the RE, TA, and TD services, respectively, were unique.

6.2. User Factors

The influence of two user characteristics (cross-platform expertise and expectations of data and function distribution across devices) on HT execution time was examined. In the pre-test questionnaire, participants answered questions to determine their level of expertise across platforms. For each question, experience points were assigned to key responses and accumulated.

A Pearson's R correlation test was carried out to investigate the relationship between user levels of cross-platform expertise and HT execution time. The correlation was significant for the RE service on both HTs (HT1: $R = -0.851, p < 0.05$; HT2: $R = -0.899, p < 0.05$). The negative R values suggest that as user cross-platform expertise increases, the time spent on completing the HT decreases. The correlation was also significant for the TA service (HT1: $R = -0.935, p < 0.05$; HT2: $R = -0.919, p < 0.05$). For the TD service, significant correlation was revealed for the first HT ($R = -0.981, p < 0.01$) and non-significant correlation for the second HT ($R = -0.734, p > 0.05$).

The second user factor considered in this study addressed users' expectations of the distributed content across devices. In the pre-test questionnaire, participants were asked a close-ended question: 'What is your expectation of content and functions of the cross-platform user interfaces?' The responses to this question represented the different levels of redundancy: exclusive, redundant, and complementary. The three services provided in the study are categorised as complementary. The findings show that participants expecting a complementary service had the lowest execution time for HTs in all three cross-platform services.

6.3. Efficiency

To measure efficiency, the HT execution time and the number of actions for each HT were collected. From these two metrics, the mean execution time (minutes) and the average number of actions were found to be interpretable. HTs performed on the TD service had the lowest average task execution time and number of actions (HT1, HT2: average time = 5.1, 3.7 min, average number of actions = 7.3, 12.2). Since the HTs across the tested services tended to have similar levels of complexity, it could be argued that the TD service supported users in performing their HTs more efficiently across platforms in comparison to the RE and TA services. RE exhibited the highest average task execution time (HT1, HT2: average time = 10.9, 7.1 min, average number of actions = 31.0, 15.4), followed closely next by TA (HT1, HT2: average time = 10.8, 12.0 min, average number of actions = 18.5, 22.5).

The study's participants attempted the HTs in six different orders of user interfaces. Since the study had a small number of participants per service, each participant attempted each HT in a distinct order. This aspect prevented us from averaging the HT execution times per order of interfaces—a type of analysis that could occur if tested on two user interfaces. With two user interfaces. Therefore, participants can only attempt each HT using one of two distinct orders. This means that it is more likely to have more participants conduct the same HTs in the same order.

To determine if we could reduce the number of cross-platform efficiency metrics, a correlation test was carried out between the metrics. The results show that the relationship between the metrics was in some cases statistically insignificant (with coefficients of less than 0.3 [67]). In general, the correlations were positive, indicating that an increase in one variable correlated with an increase in the other (RE—H1: 0.64, H2: 0.77; TA—H1: 0.21, H2: 0.96; TD—H1: 0.23, H2: 0.58). However, the correlation results were not always consistent across the data sets at the same level. If the metrics were consistently correlated in a specific pattern, we may have been able to argue that employing one metric could be adequate to assess efficiency. However, an interpretation of the correlations could be that each metric adds unique findings.

6.4. Effectiveness

Cross-platform effectiveness was considered using two metrics: task completion rate and the number of errors. RE participants generally made more mistakes than did the TA and TD participants

(HT1, HT2: completion rate = 20%, 50%, average errors = 10.8, 5.8). This result could be an indication that many cross-platform design issues need to be addressed in the RE service. TA participants performed only slightly better than the RE participants (HT1, HT2: completion rate = 25%, 50%, average errors = 3.2, 5.2). Overall, the TD service had the highest completion rate and the lowest number of errors (HT1, HT2: completion rate = 50%, 70%, average errors = 1.5, 3.0). This observation could also mean that TD supports its users to complete their tasks effectively compared to the other tested services. These examples demonstrate the usefulness of the metrics for determining cross-platform effectiveness.

For metric reduction purposes, the relationship between cross-platform effectiveness metrics were examined (the task completion rate and the number of errors) for each HT (RE—H1: −0.69, H2: −0.79; TA—H1: −0.98, H2: −0.99; TD—H1: −0.87, H2: −0.94). The correlation between the variables in all cases was significant. The correlations were negative, which can be interpreted as an increase in the number of errors that users encounter across devices that correlates with a decrease in completion rates. The correlation results appear consistent across the data sets; hence, it can be assumed that similar information will be obtained if only a single metric is used.

6.5. Continuity

The continuity factor is concerned with how seamlessly a user continues with an interrupted task on a new user interface. As previously indicated, three metrics were selected to measure the continuity factors: task-resuming success, the time taken to resume tasks, and user satisfaction with the seamlessness of the transition.

The STS was completed after participants undertook two subtasks that involved moving from one device to another (switching between two devices). Factor analysis was performed by first unifying the scales so that one represented the negative and five represented the positive for all statements. This means that the responses to Statement 2—the negative statement—were transformed to conform to the positive statements on the scales. Each participant conducted two HTs, and this resulted in four transitions per user. Factor analysis was carried out on 52 transitioning cases. Eigenvalues were calculated and used to decide how many factors should be extracted during the overall factor analysis. The eigenvalue for a given factor measures the variance in all variables that is accounted for by that specific factor. It has been suggested that only factors with eigenvalues greater than one should be retained [68]. The eigenvalue output indicates that there was only one significant factor among the three statements in the STS, the scale as a whole might reflect only one dimension: seamless transitioning. This finding could be interpreted as an indicator that the results of the participant responses to the statements on the scale could be used to measure the seamlessness of the transition between user interfaces.

The correlation between statements was also investigated on the findings of the STS based on the 52 responses. The results show that all statements correlated significantly with one another—all correlations were significant at 0.01 (2-tailed) (see Table 3). The results show a positive correlation between the statements. That is, increases in one statement correlated with increases in the others. This finding supports our assumption that the proposed scale could be used to measure only one dimension: the seamless transition between devices; that is, if the coefficient of any statement decreases (e.g., to below 0.3 [67]), it can be concluded that the statement does not contribute to the overall seamless transition score.

Table 3. Seamless Transition Scale (STS) statements correlation.

Statement	1	2	3
S1	1		
S2	0.740	1	
S3	0.787	0.613	1

Further analysis was conducted to test the reliability of the scale. A reliability test was conducted on the combined data sets (the 52 responses to STS) obtained by testing the three cross-platform services: RE, TA and TD. The results show high reliability values, with Cronbach's alpha equal to 0.882, which exceeds the acceptable reliability coefficient value of 0.7 [69]. This outcome indicates that the STS is a reliable method that can be used to measure seamless transitioning between devices.

The overall results for the task-resuming success rate (RE, TA, TD: 40%, 62.5%, 93.7%), the time taken to resume tasks (RE, TA, TD: 12.3, 12.2, 5.8 min), and the user satisfaction with the seamlessness of transitions (RE, TA, TD: 2.8/5, 2.5/5, 3.8/5) showed that the TD cross-platform service supported continuity better than the RE and TA services. However, these are the overall results for continuity, and they are not able to verify which user interface worked better to support task continuity and which failed to do so. Therefore, analysing data by participant could provide an indication of which user interface should be improved to support task continuity.

A Pearson correlation test was carried out to investigate the relationship between the continuity metrics. The test was performed on the data obtained from the first and second transitions of the first HT of each service (see Table 4 and 5, respectively). In some cases, there were significant correlations between variables. One interpretation of the results indicates that increases in the time taken to resume tasks correspond to decreases in participant satisfaction about the smoothness of transitions as well as decreases in task-resuming success (except for TD, which showed an increase in the second transition). The correlation of the results of task-resuming success and STS were mostly positive, which indicates that increases in task-resuming success correspond to increases in STS scores. As shown in two of the cases, increases in task-resuming success correlate with decreases in seamless transition, which means that users may not always be satisfied with the smoothness of transitions even when they are able to resume a task successfully. As the three metrics did not show a consistent pattern of correlation across the data sets, it could be argued that each metric provided information that was different from that of the other metrics and should therefore all be used in cross-platform usability evaluations.

Table 4. Pearson correlation between continuity metrics for first transition. M1: task-resuming success rate, M2: average time to resume task in minutes, and M3: average seamless transition.

	M3			**M1**		
	RE	**TA**	**TD**	**RE**	**TA**	**TD**
M1	−0.19	0.75	0.50			
M2	−0.63	−0.60	−0.34	−0.09	−0.96	−0.80

Table 5. Pearson correlation between continuity metrics for second transition. M1: task-resuming success rate, M2: average time to resume task in minutes, and M3: average seamless transition.

	M3			**M1**		
	RE	**TA**	**TD**	**RE**	**TA**	**TD**
M1	−0.76	0.85	0.57			
M2	−0.78	−0.82	−0.84	−0.98	−0.42	−0.38

6.6. Productivity

The unproductive period in minutes for each HT was computed and collected in the study (RE—H1: 6.03, H2: 4.08; TA—H1: 2.59, H2: 5.58; TD—H1: 2.23, H2: 4.64). The results show that RE's first HT and TA's second HT produced the highest average of unproductive periods. The majority of participants of these two HTs made several mistakes and requested help when attempting subtasks across devices. For instance, during RE's first HT, participants encountered inconsistency issues with the search panel across the difference user interfaces. It is safe to conclude that the unproductive period metric can contribute new information about cross-platform usability since it takes into account the time spent seeking assistance.

6.7. Satisfaction

The CPUS was designed to produce a single reference score of user satisfaction for a cross-platform service. The analysis of the CPUS results began with the inspection of the statements on the scale separately. Table 6 lists descriptive statistics of CPUS responses for the three test services. The table shows the actual mean, post-transformation (PT) mean, and the standard deviation (SD) of the CPUS responses for each of the eight statements. The PT mean adjusts the scores of negative statements 4, 6, 7 and 8, so that positive responses are associated with a larger number, as are the other four (positive) statements. In the PT mean column, the larger numbers for positive statements 1, 2, 3 and 5 show that these statements received more positive responses (agreements), and the larger numbers for negative statements 4, 6, 7 and 8 indicate that these received more positive responses (disagreements). In the RE service, statement 8 received the largest number of positive responses (disagreements) with a mean of 2.20. For the TA service, statements 1, 2, 3, 4, 5 and 6 had the largest positive responses equally. For the TD service, participants gave statement 1 the largest number of positive responses with an average of 4.00, which was also the largest number across all statements for all tested cross-platform services. These findings conclude that the scales can generate different scores according to the designs of services. The results in Table 6 are also consistent with other cross-platform usability measures. For instance, the TD service received more positive impressions on the scale compared to the RE and TA services. This result is consistent with the cross-platform execution times, the completion rate, and the continuity results. TD users performed HTs with shorter execution times, had greater task completion rates, and produced better continuity results compared with RE and TA users.

Table 6. Mean, post-transformation mean (PTM), and standard deviation (SD) of Cross-Platform Usability Scale (CPUS) responses.

Statement	RE			TA			TD		
	Mean	PTM	SD	Mean	PTM	SD	Mean	PTM	SD
S1	1.80	1.80	0.83	2.25	2.25	1.25	4.00	4.00	0.81
S2	1.80	1.80	0.83	2.25	2.25	1.25	3.50	3.50	1.29
S3	1.80	1.80	0.83	2.25	2.25	1.25	3.50	3.50	1.29
S4	4.20	1.80	0.83	3.75	2.25	1.25	2.75	3.25	0.95
S5	1.80	1.80	0.83	2.25	2.25	1.25	3.25	3.25	0.95
S6	4.40	1.60	0.89	4.00	2.00	1.41	2.50	3.50	1.29
S7	4.40	1.60	0.89	3.75	2.25	1.25	2.50	3.50	1.29
S8	3.80	2.20	0.44	4.00	2.00	1.41	2.50	3.50	1.29

The correlation between CPUS statements was investigated by carrying out a correlation analysis on the CPUS results from the three test services. The Pearson correlation results shown in Table 7 confirm that the eight statements correlate significantly with one another (with coefficients greater than 0.3 [67]). All statements exhibited significant coefficient values that indicate that each statement contributes positively to measuring the overall cross-platform satisfaction.

Table 7. CPUS statement correlation.

Statement	1	2	3	4	5	6	7	8
S1	1							
S2	0.63	1						
S3	0.63	1.00	1					
S4	0.42	0.94	0.94	1				
S5	0.42	0.94	0.94	1.00	1			
S6	0.63	1.00	1.00	0.94	0.94	1		
S7	0.63	1.00	1.00	0.94	0.94	1.00	1	
S8	0.63	1.00	1.00	0.94	0.94	1.00	1.00	1

A factor analysis of the CPUS statements was performed to confirm that the statements addressed different dimensions of the participants' experience. Figure 4 illustrates the eigenvalues output of the factor analysis of the test services' generated data sets, as well as the results of the analysis of the data set combination. The results show that the first statement was the only significant factor among the eight statements. It can be concluded that viewed collectively, the statement reflects users' satisfaction with cross-platform systems.

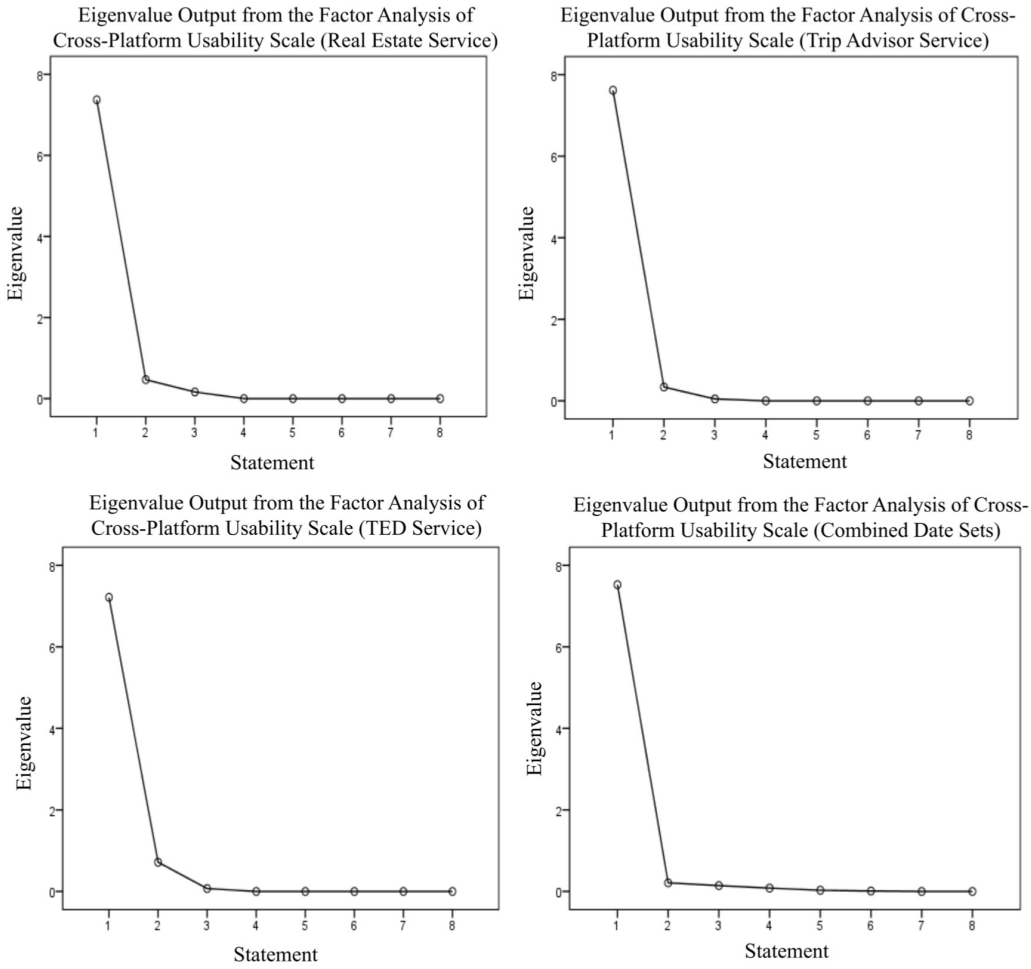

Figure 4. Factor analysis Eigenvalue output from the CPUS responses.

The internal consistency of the CPUS statement was tested via a reliability test. The PT mean was used to compute this test (RE: 0.98; TA: 0.99; TD: 0.98; Combined: 0.99). The results of the reliability test, where the high Cronbach's alpha corroborates the viability and reliability of the CPUS as a measure for cross-platform satisfaction [69].

7. Discussion and Model Refinement

7.1. Data Collection Techniques

The combination of data collection methods in cross-platform assessment model supported the identification of several usability issues. Of the three methods, the think-aloud protocol generated the most cross-platform usability issues for all three test services which is consistent with Nielsen's claim [11]. Nevertheless, our findings highlight the importance of the three methods to ensure the identification of as many usability issues as possible that might not overlap across methods.

7.2. User Factors

Usability measures are affected by the context of use, including user and task characteristics [53]. That is, the effectiveness of a usability test depends on the given tasks, the methodology, and the users' characteristics [70]. In the design of this study, the tasks were formulated in a way that reduced the effects of task characteristics on cross-platform usability measures. Nonetheless, user traits should be inspected as they may affect the cross-platform usability test. The study findings show that user characteristics, such as cross-platform expertise and distributed data and functions expectation, could have an impact on HT execution time. Accordingly, user factors should be considered when evaluating the usability of cross-platform services. This consideration entails the careful recruitment of participants to eliminate the likelihood of their impact on overall measurements.

7.3. Correlation between Metrics

Several metrics were used to assess efficiency, effectiveness, and continuity. Correlation tests were carried out to determine their impact. For the efficiency factors, the findings show primarily non-significant correlation between the execution time and the number of actions. These results confirm the values of each metric and their ability to contribute unique findings. In the case of the effectiveness factors, the correlation between the metrics was found to be significant, and hence, the use of a single metric can potentially suffice. Nevertheless, there is still a risk of overlooking valuable issues when discarding one of two effectiveness measures. For instance, users are capable of completing a task successfully while still encountering difficulties and making mistakes. In the presented cross-platform usability study, participants incurred several errors due to their attempts at reusing prior knowledge from one device to inform their interaction with another. This observation shows that counting the number of errors can indicate a number of design issues in a cross-platform service that would have otherwise been overlooked. Similar to efficiency, the continuity metrics correlation results proved statistically insignificant, which informs their value at identifying unique usability measures.

7.4. Metrics Comprehensiveness

The proposed cross-platform usability model utilised both traditional usability metrics and introduced a new metric—'continuity'. All metrics proved valuable; however, it is difficult to argue for the comprehensiveness of these metrics for a cross-platform usability assessment specifically. This difficulty is largely because the theme of cross-platform usability is very much in its infancy and requires further studies to understand its complexities. In relation to the proposed model, we believe that further refinement is expected to confirm the relevance of the CPUS and STS statements. Additionally, further studies may aid the expansion of the model by considering more factors, such as learnability.

7.5. The Number and Order of User Interfaces

The study assessed the usability of three user interfaces, which proved time consuming and generated fewer interpretable results. For instance, it proved difficult to extract measurements (e.g., average HT execution time) because each participant attempted each HT in a distinct order due to the larger number of combination options.

Observations show that at times it was difficult for some participants to explain inter-usability problems because they were exposed to three different user interfaces. Furthermore, it was noted that experience can be gained as they interacted with one interface and then another. This finding further emphasises the importance of testing user interfaces in pairs.

Arguably, users will still need to interact with user interfaces from different platforms interchangeably. In fact, it was observed in the study that unique inter-usability problems may occur when participants interacted with the user interfaces in a specific order. These problems are due to the influence of past experience with the previous user interfaces in the HT assignment.

Therefore, testing the interchangeability of user interfaces remains important in cross-platform usability evaluations.

7.6. Model Refinement

Based on the study findings, the proposed model was refined as shown in Figure 5. The refined model maintains the combination of data collection methods and cross-platform usability metrics from the original model as they have been found to be invaluable to investigating the unique behaviours exhibited in cross-platform systems. The previous section highlighted the importance of user interface interchangeability as unique cross-platform issues can be identified in each distinctive order. Nonetheless, the findings support testing with user interface pairs, particularly when working with a small sample group. In the refined model, two new factors are included: cross-platform expertise and user expectations of data and function distribution. These factors should be taken into account when screening participant for cross-platform testing.

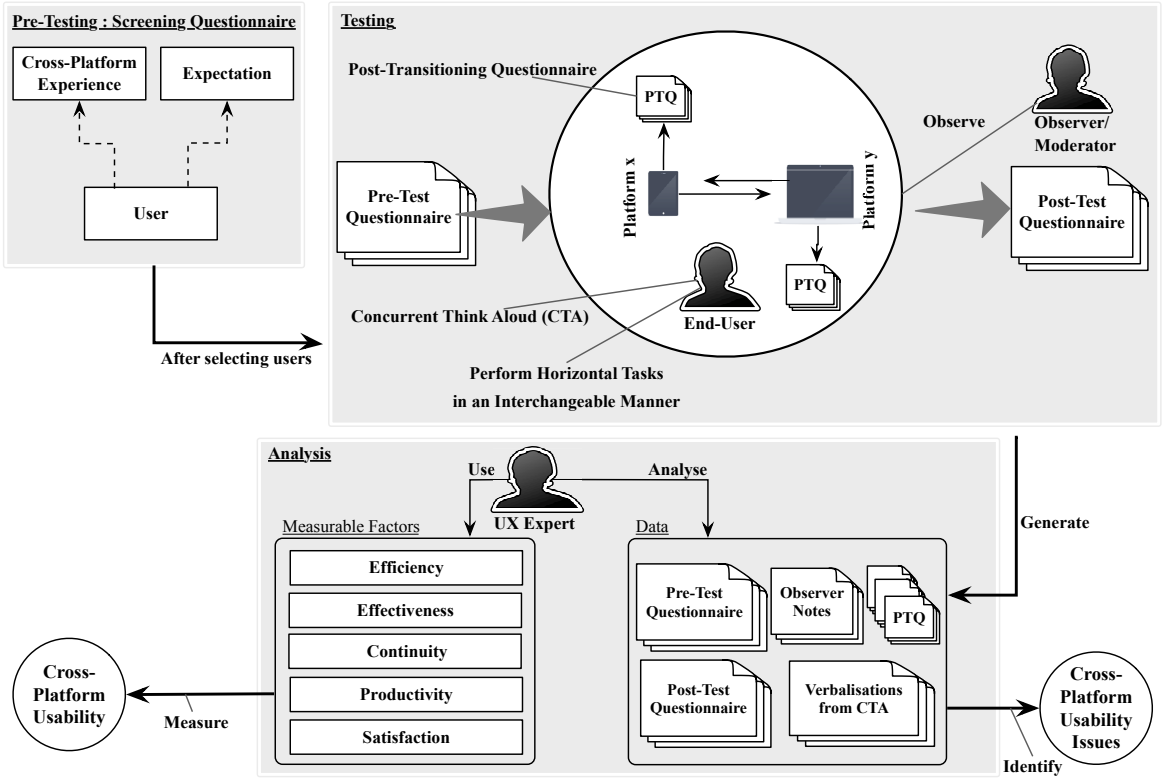

Figure 5. The refined cross-platform usability assessment model.

8. Expert Evaluation of the Model

We carried out an expert evaluation to capture the opinions and impressions of professionals specialising in UX about the model. We designed a questionnaire that involved a part that captures participant demographics and their roles in their respective organisations, and another part to capture participant opinions about the model. The questionnaire also involved detailed explanations of our assessment model. Participant responses were analysed using predefined themes including model appropriateness, effectiveness and usefulness, model completeness and redundancy, adaptation of the model to practice and challenges.

Participants' most recent study programmes included computer science, software engineering, information systems, HCI, and business. The age of participants ranged from 30 to 39 years old for six participants, 40 to 49 for one participant and 50 years old or more for one participant. Three participants held master degrees, and five participants held bachelor degrees. Of the participants,

there were two UX designers, two usability engineers, one information architect, one interaction designer, one usability analyst and one UX specialist. Three participants had been working in the field of UX or usability for more than ten years and five participants for between five and ten years. The categories that describe the industry that the participants primarily worked in ranged across government, education, media, medicine and software. Four participants described UX as a part of their roles in their organisations, two participants mentioned that they worked as a part of a UX team in their organisations and two participants defined themselves as the UX people in their organisations.

Generally, seven participants agreed that the assessment model could be used to address some of the current gaps in cross-platform usability evaluation approaches. One participant, P1, stated that they felt that the model is 'appropriate' and its elements are designed carefully to capture information about cross-platform usability. Another participant, P3, mentioned that the model could contribute to the cross-platform service design practices and may open the door to new thinking about user-centred service design. One participant, P8, focused more on the appropriateness of the model from the business perspective, rather than its suitability for assessing cross-platform usability. This participant argued that the appropriateness of the model needed to be determined based on the practical return on investment, which was not clear in the description of the model.

In terms of the effectiveness of the model for identifying cross-platform usability problems, all participants (except participant P8) agreed that using the model could help identify cross-platform usability problems. One participant, P1, linked the model's effectiveness to its focus on cross-platform usability assessment. That is, the model does not integrate cross-platform usability evaluation procedures with the traditional procedures. The participant considered this as an important advantage of the model, since it could allow evaluators to focus on identifying cross-platform usability problems. Participant, P3, saw that the model might be effective for identifying inter-usability problems because of our approach to designing HTs. That is, our way of designing HTs allows users to take the same or a similar path to achieve subtasks across devices. Hence, when switching from one device to another, evaluators could identify problems by observing participant interactions and behaviours when attempting to reuse knowledge of specific functions. Participant, P4, reported that the use of several data-collection techniques in our model could help to identify the maximum potential cross-platform usability problems. However, participant, P8, suggested that more practical examples may be needed to allow for the assessment of the effectiveness of the model.

Concerning the usefulness of the model, on a Likert scale from one to five, where one is not useful, three is neutral and five is extremely useful, five participants responded that the model is extremely useful, two participants said that it is useful and one participant remained neutral. Six participants also mentioned that they would use the model in cases when they needed to assess the cross-platform usability of a service and one participant indicated that he might use it by focusing only on certain measures. Participant, P8, stated that he might not use the model because it lacked a description of its business value, which would be needed to fund the assessment activity.

In regard to the completeness of the model, seven participants who answered the related question agreed that the model is comprehensive. One participant stated: "the model is comprehensive, where all required data-collection techniques that are normally used in usability testing are included. Also, it employs most of the required measurements to support the evaluations" (P1; UX designer in the education domain). In terms of the redundancy of model elements (e.g., measures), participants stated that there is no redundancy in the model.

Our analysis of all participants' responses to the question "Do you think that the model can be adopted in the UX evaluation practices? Why?" showed that seven participants generally agreed that the model can be adopted in the UX evaluation practices. One participant states, "Yes, this can be for several reasons: the absences of a current method for cross-platform usability testing, the simplicity and consistency of the model with traditional testing methods and the revolution of multi-platform services." (P1; UX designers in the education domain).

In regard to challenges and constraints for adopting the model, three participants, P4, P6 and P7, indicated that there are no challenges or constraints for adopting the model for UX evaluation practice. However, some participants listed some possible challenges for the quick adoption of the model. For example, three participants, P1, P2 and P5 mentioned that usability evaluators might need to understand the cross-platform usability concepts and be trained about how to use the model to adopt it and execute it effectively. Participant, P3, indicated that there might be challenges integrating the model into current user-centred design (UCD) practices, concerning the design for individual user interfaces, rather than MUIs. Participant, P8, indicated that there might be a need for a description of the business value of the project, which could be required to fund the assessment activity so businesses could determine whether to adopt the model or not.

9. Conclusions and Future Work

This paper presented a cross-platform usability assessment model and conducted a usability assessment for three cross-platform services using the proposed model. The findings have shown that the think-aloud protocol is the most valuable method for revealing cross-platform usability issues; however, other methods such as observation also help to uncover usability issues. Furthermore, the results have shown that differences in user levels of cross-platform expertise and the expectation of levels of data and function redundancy across devices influence cross-platform task execution times. The study demonstrated that the results obtained from the usability metrics proposed in this model are valuable. The results indicate that the statements comprising each scale (cross-platform usability and seamless transitioning) correlated significantly, that they were internally consistent and that they addressed only one dimension.

One of the most important strengths of the model is that it includes objective and subjective measures for evaluating the continuity and the seamless transition between devices, which are the most important aspects of inter-usability. The CPUS can also reveal important information about the inter-usability of a cross-platform service (e.g., inter-platform service consistency), making it a valuable tool that can be used in cross-platform user experience studies.

It is important to mention that the analysis of the cross-platform usability of a combination of three user interfaces was found to be time consuming and the results were less interpretable. Therefore, we recommend testing user interfaces in pairs. Testing user interfaces in an interchangeable manner is also an important aspect of cross-platform usability evaluation since unique issues can be identified for each distinctive order.

The proposed model was refined based on the user study findings and discussion. In addition, eight UX experts evaluated our model and most of them agreed that the model is appropriate, useful and can be adopted in the practice.

In future work, we intend to improve and extend the user study and assessment model and examine various factors, in the following ways:

- Confirm the viability of the refined cross-platform assessment model with a larger number of participants.
- Conduct more experimental studies to determine whether the proposed assessment model could be employed for testing services with different degrees of data and functions of redundancy (e.g., exclusive), as well as for testing services when users interact with them to achieve unrelated tasks in a sequential mode, and when attempting related and unrelated tasks in a simultaneous mode.
- Utilise other approaches (e.g., using severity ratings or classification of problems) in addition to the problem counting approach used in the study, as this technique may have some limitations (e.g., the counts of potential problems may include problems that are not real usability problems) [71].
- Address the limitation of having only one evaluator to analyse the cross-platform usability problem by recruiting multiple evaluators.

- Compare the proposed assessment model to other usability evaluation methods that can be used to assess cross-platform usability (e.g., expert review). This would assist, for example, in choosing the superior method for finding serious cross-platform usability problems with the least amount of effort.
- Conduct further studies to extend the model by considering more factors, such as learnability.

Author Contributions: Conceptualization, K.M., M.H. and A.L.U.; Formal analysis, K.M.; Investigation, K.M.; Methodology, K.M.; Project administration, K.M.; Resources, K.M., M.H. and A.L.U.; Supervision, M.H. and A.L.U.; Validation, K.M., M.H., A.L.U. and S.A.-M.; Visualization, K.M. and S.A.-M.; Writing—original draft, K.M. and S.A.-M.; Writing—review & editing, S.A.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to extend their gratitude and acknowledgements to all study participants for their time and energy spent on this project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Harper, R.; Rodden, T.; Rogers, Y.; Sellen, A. *Being Human: Human-Computer Interaction in the Year 2020*; Microsoft Research: Cambridge, UK, 2008.
2. Rogers, Y. The changing face of human-computer interaction in the age of ubiquitous computing. In *Symposium of the Austrian HCI and Usability Engineering Group*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–19.
3. Seffah, A.; Forbrig, P.; Javahery, H. Multi-devices "Multiple" user interfaces: Development models and research opportunities. *J. Syst. Softw.* **2004**, *73*, 287–300. [CrossRef]
4. Think with Google. The New Multi-Screen World: Understanding Cross-Platform Consumer Behavior. 2012. Available online: https://www.thinkwithgoogle.com/advertising-channels/mobile-marketing/the-new-multi-screen-world-study/ (accessed on 2 April 2019).
5. Brudy, F.; Holz, C.; Rädle, R.; Wu, C.J.; Houben, S.; Klokmose, C.N.; Marquardt, N. Cross-Device Taxonomy: Survey, Opportunities and Challenges of Interactions Spanning Across Multiple Devices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems 2019*; Association for Computing Machinery (ACM): New York, NY, USA, 2019.
6. Weiser, M. The computer for the 21st century. *IEEE Pervasive Comput.* **2002**, *1*, 19–25. [CrossRef]
7. Jokela, T.; Ojala, J.; Olsson, T. A diary study on combining multiple information devices in everyday activities and tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2015; pp. 3903–3912.
8. Forrester Research. The European tablet landscape. A Technographics Data Essentials Document. 2014. Available online: https://www.forrester.com/report/The+European+Tablet+Landscape/-/E-RES91561 (accessed on 2 April 2019).
9. Santosa, S.; Wigdor, D. A field study of multi-device workflows in distributed workspaces. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*; ACM: New York, NY, USA, 2013; pp. 63–72.
10. Scharf, F.; Wolters, C.; Herczeg, M.; Cassens, J. Cross-Device Interaction Definition, Taxonomy and Applications. In Proceedings of the Third International Conference on Ambient Computing, Applications, Services and Technologies (IARIA), Porto, Portugal, 29 September–3 October 2013.
11. Nielsen, J. Usability inspection methods. In *Conference Companion on Human Factors in Computing Systems*; ACM: New York, NY, USA, 1994; pp. 413–414.
12. Seffah, A.; Donyaee, M.; Kline, R.B.; Padda, H.K. Usability measurement and metrics: A consolidated model. *Softw. Qual. J.* **2006**, *14*, 159–178. [CrossRef]
13. Salvendy, G. *Handbook of Human Factors and Ergonomics*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
14. Albert, W.; Tullis, T. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*; Morgan Kaufmann: Burlington, MA, USA, 2013.

15. Unterkalmsteiner, M.; Gorschek, T.; Islam, A.M.; Cheng, C.K.; Permadi, R.B.; Feldt, R. Evaluation and measurement of software process improvement?a systematic literature review. *IEEE Trans. Softw. Eng.* **2012**, *38*, 398–424. [CrossRef]
16. Lewis, J.R. Usability: Lessons learned? and yet to be learned. *Int. J. -Hum.-Comput. Interact.* **2014**, *30*, 663–684. [CrossRef]
17. Denis, C.; Karsenty, L. Inter-usability of multi-device systems: A conceptual framework. In *Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces;* Wiley: Hoboken, NJ, USA, 2004; pp. 373–384.
18. Majrashi, K.; Hamilton, M. A cross-platform usability measurement model. *Lect. Notes Softw. Eng.* **2015**, *3*, 132. [CrossRef]
19. Majrashi, K.; Hamilton, M.; Uitdenbogerd, A.L. Correlating cross-platform usability problems with eye tracking patterns. In *Proceedings of the 30th International BCS Human Computer Interaction Conference: Fusion*; BCS Learning & Development Ltd.: Swindon, UK, 2016; p. 40.
20. Majrashi, K.; Hamilton, M.; Uitdenbogerd, A.L. Cross-platform cross-cultural user experience. In *Proceedings of the 30th International BCS Human Computer Interaction Conference: Fusion*; BCS Learning & Development Ltd.: Swindon, UK, 2016; p. 20.
21. Dong, T.; Churchill, E.F.; Nichols, J. Understanding the challenges of designing and developing multi-device experiences. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*; ACM: New York, NY, USA, 2016; pp. 62–72.
22. Shin, D.H. Cross-platform users? Experiences toward designing interusable systems. *Int. J. Hum.-Comput. Interact.* **2016**, *32*, 503–514. [CrossRef]
23. Rieger, C.; Majchrzak, T.A. Towards the Definitive Evaluation Framework for Cross-Platform App Development Approaches. *J. Syst. Softw.* **2019**, *153*, 175–199. [CrossRef]
24. Antila, V.; Lui, A. Challenges in designing inter-usable systems. In *IFIP Conference on Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 396–403.
25. Samaan, K.; Tarpin-Bernard, F. Task models and interaction models in a multiple user interfaces generation process. In *Proceedings of the 3rd Annual Conference on Task Models and Diagrams*; ACM: New York, NY, USA, 2004; pp. 137–144.
26. Seffah, A.; Javahery, H. *Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2004; pp. 11–26.
27. Pyla, P.S.; Tungare, M.; Pérez-Quinones, M. Multiple user interfaces: Why consistency is not everything, and seamless task migration is key. In Proceedings of the CHI 2006 Workshop on the Many Faces of Consistency in Cross-Platform Design, Montreal, QC, Canada, 22–23 April 2006.
28. Nilsson, L. *Continuity of Service in Design for a Specific Platform: Combining Service-and Interaction Design Perspectives in a Multiple Platform Environment*; Institutionen for Datavetenskap: Umea, Sweden, 2006.
29. Ali, M.F.; Perez-Quinones, M.A.; Abrams, M.; Shell, E. Building multi-platform user interfaces with UIML. In *Computer-Aided Design of User Interfaces III*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 255–266.
30. Meskens, J.; Vermeulen, J.; Luyten, K.; Coninx, K. Gummy for multi-platform user interface designs: Shape me, multiply me, fix me, use me. In *Proceedings of the Working Conference on Advanced Visual Interfaces*; ACM: New York, NY, USA, 2008; pp. 233–240.
31. Bång, M.; Larsson, A.; Berglund, E.; Eriksson, H. Distributed user interfaces for clinical ubiquitous computing applications. *Int. J. Med. Inform.* **2005**, *74*, 545–551. [CrossRef]
32. Tesoriero, R.; Lozano, M.; Vanderdonckt, J.; Gallud, J.A.; Penichet, V.M. Distributed user interfaces: Collaboration and usability. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2012; pp. 2719–2722.
33. Bouabid, A.; Lepreux, S.; Kolski, C. Design and evaluation of distributed user interfaces between tangible tabletops. *Univ. Access Inf. Soc.* **2019**, *18*, 801–819. [CrossRef]
34. Segerståhl, K. Utilization of pervasive IT compromised? Understanding the adoption and use of a cross media system. In *Proceedings of the 7th International Conference on Mobile and Ubiquitous Multimedia*; ACM: New York, NY, USA, 2008; pp. 168–175.
35. Wäljas, M.; Segerståhl, K.; Väänänen-Vainio-Mattila, K.; Oinas-Kukkonen, H. Cross-platform service user experience: A field study and an initial framework. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*; ACM: New York, NY, USA, 2010; pp. 219–228.

36. Levin, M. *Designing Multi-Device Experiences: An Ecosystem Approach to User Experiences Across Devices*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2014.
37. O'Leary, K.; Dong, T.; Haines, J.K.; Gilbert, M.; Churchill, E.F.; Nichols, J. The moving context kit: Designing for context shifts in multi-device experiences. In *Proceedings of the 2017 Conference on Designing Interactive Systems*; ACM: New York, NY, USA, 2017; pp. 309–320.
38. Sánchez-Adame, L.M.; Mendoza, S.; Viveros, A.M.; Rodríguez, J. Towards a Set of Design Guidelines for Multi-device Experience. In *International Conference on Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 210–223.
39. Biørn-Hansen, A.; Grønli, T.M.; Ghinea, G. A survey and taxonomy of core concepts and research challenges in cross-platform mobile development. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–34. [CrossRef]
40. Rieger, C.; Kuchen, H. A model-driven cross-platform app development process for heterogeneous device classes. In Proceedings of the 52nd Hawaii International Conference on System Sciences, Maui, HI, USA, 8–11 January 2019.
41. Dearman, D.; Pierce, J.S. It's on my other computer: Computing with multiple devices. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*; ACM: New York, NY, USA, 2008; pp. 767–776.
42. Bandelloni, R.; Paternò, F. Flexible interface migration. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*; ACM: New York, NY, USA, 2004; pp. 148–155.
43. Chu, H.h.; Song, H.; Wong, C.; Kurakake, S.; Katagiri, M. Roam, a seamless application framework. *J. Syst. Softw.* **2004**, *69*, 209–226. [CrossRef]
44. Kjeldskov, J.; Skov, M.B. Creating realistic laboratory settings: Comparative studies of three think-aloud usability evaluations of a mobile system. In Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction, Zurich, Switzerland, 1–5 September 2003; pp. 663–670.
45. Alhadreti, O.; Mayhew, P. Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2018; p. 44.
46. Lazar, J.; Feng, J.H.; Hochheiser, H. *Research Methods in Human-Computer Interaction*; Morgan Kaufmann: Burlington, MA, USA, 2017.
47. Holzinger, A. Usability engineering methods for software developers. *Commun. ACM* **2005**, *48*, 71–74. [CrossRef]
48. Sauro, J.; Dumas, J.S. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2009; pp. 1599–1608.
49. Brooke, J. SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
50. Albert, W.; Tullis, T.; Tedesco, D. In *Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies*; Morgan Kaufmann: Burlington, MA, USA, 2009.
51. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **1932**, *22*, 55.
52. Bevan, N. International standards for HCI and usability. *Int. J. Hum. Comput. Stud.* **2001**, *55*, 533–552. [CrossRef]
53. Bevan, N. Measuring usability as quality of use. *Softw. Qual. J.* **1995**, *4*, 115–130. [CrossRef]
54. Pyla, P.S.; Tungare, M.; Holman, J.; Pérez-Quiñones, M.A. Continuous user interfaces for seamless task migration. In *International Conference on Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 77–85.
55. Nielsen, J.; Landauer, T.K. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 1993; pp. 206–213.
56. Spool, J.; Schroeder, W. Testing web sites: Five users is nowhere near enough. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2001; pp. 285–286.
57. Faulkner, L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behav. Res. Methods Instrum. Comput.* **2003**, *35*, 379–383. [CrossRef]
58. Turner, C.W.; Lewis, J.R.; Nielsen, J. Determining usability test sample size. *Int. Encycl. Ergon. Hum. Factors* **2006**, *3*, 3084–3088.
59. Hwang, W.; Salvendy, G. Number of people required for usability evaluation: The 10 ± 2 rule. *Commun. ACM* **2010**, *53*, 130–133. [CrossRef]

60. Schmettow, M. Sample size in usability studies. *Commun. ACM* **2012**, *55*, 64–70. [CrossRef]
61. Cazañas, A.; de San Miguel, A.; Parra, E. Estimating sample size for usability testing. *Enfoque UTE* **2017**, *7*, 172–185. [CrossRef]
62. Van den Haak, M.J.; De Jong, M.D. Exploring two methods of usability testing: Concurrent versus retrospective think-aloud protocols. In Proceedings of the IEEE International Professional Communication Conference (IPCC 2003), Orlando, FL, USA , 21–24 September 2003; p. 3.
63. Jeffries, R.; Miller, J.R.; Wharton, C.; Uyeda, K. User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 1991; Volume 91, pp. 119–124.
64. Nielsen, J. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 1992; pp. 373–380.
65. Mankoff, J.; Dey, A.K.; Hsieh, G.; Kientz, J.; Lederer, S.; Ames, M. Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2003; pp. 169–176.
66. Markopoulos, P.; Bekker, M. On the assessment of usability testing methods for children. *Interact. Comput.* **2003**, *15*, 227–243. [CrossRef]
67. Hair, J.F. *Multivariate Data Analysis*; Pearson Education India: Chennai, India, 2006.
68. Kaiser, H.F. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* **1960**, *20*, 141–151. [CrossRef]
69. Nunnally, J.C.; Bernstein, I.H.; Berge, J.M.T. *Psychometric Theory*; McGraw-Hill: New York, NY, USA, 1967; Volume 226.
70. Molich, R.; Ede, M.R.; Kaasgaard, K.; Karyukin, B. Comparative usability evaluation. *Behav. Inf. Technol.* **2004**, *23*, 65–74. [CrossRef]
71. Hornbæk, K. Dogmas in the assessment of usability evaluation methods. *Behav. Inf. Technol.* **2010**, *29*, 97–111. [CrossRef]