BILINGUAL PUBLISHING GROUP

Journal of Computer Science Research

https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Expert Review on Usefulness of an Integrated Checklist-based Mobile Usability Evaluation Framework

Hazura Zulzalil[1*]**, *Hazwani Rahmat***[2]**, *Abdul Azim Abd Ghani***[1]**, *Azrina Kamaruddin***[1]

[1] *Department of Software Engineering and Information System, Universiti Putra Malaysia, UPM Serdang, Selangor, 43400, Malaysia*

[2] *Department of Information Technology, Centre for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Pagoh Higher Education Hub, Pagoh, Johor, 84600, Malaysia*

ABSTRACT

Previous mobile usability studies are only pertinent in the context of ergonomics, physical user interface, and mobility aspects. In addition, much of the previous mobile usability conception was built on desktop computing measurements, such as desktop and web application checklists, or scarcely addressed the mobile user interface. Moreover, the studies focus mainly on interface features for desktop applications and do not reflect comprehensive mobile interface features such as navigation drawers and spinners. Therefore, conducting usability evaluation using conventional usability measurement would result in irrelevant results. In addition, the resulting works are tailored for usability testing, which requires highly skilled evaluators and usability specialists (e.g., usability testers and user experience designers), who are rarely integrated into a development team. The lack of expertise could lead to unreliable usability evaluations. This paper presents a review from industrial experts on a comprehensive and feasible usability evaluation framework developed in our previous work. The framework is dedicated to smartphone apps, which integrate evaluator skills and design concerns. However, there is no evidence of its usefulness in practice. Therefore, the usefulness of the framework measurement for evaluating apps' usability in the eyes of non-usability specialists is empirically assessed in this paper through an expert review. The expert review involved eleven industrial developers and was complemented by a semi-structured interview. The method is replicated in comparison with a framework from another study. The findings show that the formulated framework significantly outperformed the framework ($p = 0.0286$) from other studies with large effect sizes ($r = 1.81$) in terms of usefulness.

Keywords: Usability framework; Mobile usability; Usability evaluation; Expert review; Heuristic walkthrough

*CORRESPONDING AUTHOR:

Hazura Zulzalil, Department of Software Engineering and Information System, Universiti Putra Malaysia, UPM Serdang, Selangor, 43400, Malaysia; Email: hazura@upm.edu.my

ARTICLE INFO

Received: 28 June 2023 | Revised: 24 July 2023 | Accepted: 26 July 2023 | Published Online: 31 July 2023

DOI: https://doi.org/10.30564/jcsr.v5i3.5816

CITATION

Zulzalil, H., Rahmat, H., Ghani, A.A.A., et al., 2023. Expert Review on Usefulness of an Integrated Checklist-based Mobile Usability Evaluation Framework. Journal of Computer Science Research. 5(3): 57-73. DOI: https://doi.org/10.30564/jcsr.v5i3.5816

1. Introduction

The user interface serves as a conduit between human and computer interactions. The evolution of the user interface has progressed from the command line interface (CLI) of a console to the graphical user interface (GUI), and later to the web user interface and mobile user interface. The evolution of the user interface has characterized usability dimensions differently for mobile applications (apps). CLI requires high memorability for competence and knowledge of entering massive commands. On the other hand, a GUI adopts a graphical representation for user interaction. Thus, learnability, effectiveness, and efficiency come first in a usable application. The Web user interface operates mostly on hypertext and multimedia elements, forming the navigational system and interconnected content. Consequently, consistency, simplicity, and information architecture play an important role in the effectiveness, efficiency, and navigability of web applications.

Likewise, a mobile user interface is shaped by its technological features. Physical device features and limitations, an integrated sensor such as proximity, tactile, or image recognition sensor, and its context of use has introduced emergent usability properties in characterizing the mobile usability dimension. Additionally, the unique data entry model, such as the use of a stylus, gestures, and the virtual keyboard, has taken place as the input device instead of a mouse and keyboard.

Mobile applications (apps) are used on-the-go, thus opening them to divided attention while used in different mobility conditions (e.g., sitting, walking, and driving). Apps offer support for a broad range of tasks (e.g., streaming online movies, browsing information, and performing online transactions) without the need for a computer. However, as mobile operating systems advanced, the user interface of mobile applications was rapidly enhanced through software updates. The update involves logical user interfaces (LUI) and graphical user interfaces (GUI), which affect apps' usability, rather than physical user interfaces (PUI), which affect the device's usability. New features and functionalities are introduced to enhance the mobile user interface with constant version updates. Interface features such as the navigation drawer and expansion panel are used to maximize the limited screen size. Meanwhile, snack boxes and toasts are used to deliver a prompt visual response when handling divided attention and mobility conditions, which could result in an accidental activation. Consequently, usability criteria such as connectivity, relevance, and responsiveness are the highlights of conceptualizing app usability. However, the smartphone is used by users of all ages with various levels of computing background. Thus, usability criteria such as familiarity, flexibility, and appropriateness are also highlighted in denoting mobile usability.

The conceptualization of the mobile usability dimension has been widely studied [1,2]. Numerous attempts have been made to characterize usability in view of performance-based measures [3], the physical user interface [4,5], mobile device concerns [6], usability principles [7], usability criteria [8], and interface features [9-12].

This study presents a measurement for evaluating the usability of mobile applications in the context of integrated evaluator skills. The measurements are developed by capturing the interface features, usability criteria, and design pattern, which augment the evaluation basis from multiple evaluators' viewpoints. The integration involves a comprehensive bridging of the semantic gap between different abstraction levels of usability constructs; interface features, usability features, and usability criteria into one integrated framework.

The remainder of this article is divided into eight sections. Section 2 reviews the existing mobile usability evaluation framework. Section 3 describes how the framework measurement is formulated. The resulting framework is presented in Section 4. The evaluation of the framework's usefulness is described in Section 5. Meanwhile, the results and discussions, and threats to validity are discussed in Sections 6 and 7. Finally, in Section 8, conclusions and future studies are presented.

2. Related works

Mobile usability evaluation methods such as field testing and lab experiments have been introduced for evaluating mobile applications. However, the limitations and difficulties resulting from these methods favor traditional usability evaluation methods such as usability testing and inspection methods. Inspection methods such as heuristic evaluation (HE) gain wide acceptance in industrial practice due to their simplicity, low cost, and short time, with no additional equipment required. Hence, a wide variety of heuristic evaluation methods besides Nielsen and Molich's ten heuristics have been developed, such as checklist-based heuristics [6]. Consequently, the use of a checklist was extended to frameworks in an effort to characterize usability. This has benefits for usability conception, design, and evaluation purposes.

This section discusses the checklist-based frameworks in four categories. The categories are attribute-based frameworks, integrated-based frameworks, theoretical-based frameworks, and decision-based frameworks. The first and third categories serve to conceptualize the usability dimension. Meanwhile, the second approach focuses on the interface feature, and the last category is specific to decision-making and prioritizing usability constructs. The framework's structural base, evaluator viewpoint, intended platform, and scope of evaluation item for each framework concerning the aim of each literature's work were compared.

2.1 Attribute-based frameworks

The increasing capabilities of mobile phones have encouraged several usability investigations to characterize the usability dimension of mobile phones. Initially, as smartphones started to emerge in 2009, Hussain and Kutar adopted a Goal Question Metric (GQM) approach for their framework in conceptualizing usability dimensions [3]. Based on ISO 9241-11 usability criteria (effectiveness, efficiency, and satisfaction) as usability parameters for the goal, a set of questions was associated with each goal, in a checklist form. The questions were further used to derive metrics as a performance indicator for each goal and question. However, since mobile-specific interface features for smartphones were just being launched at the time, the outcome focuses more on the logical user interface.

Further, another study by Saleh et al. [8] adopted the same approach, GQM, in constructing their framework. In contrast to Hussain and Kutar's approach [3], they developed a more comprehensive set of usability criteria denoting mobile applications (apps) by extending the PACMAD model [5] as the base of their framework structure.

Though both studies managed to conceptualize usability, the use of the GQM approach resulted in a metrics-oriented performance-based checklist that scarcely acknowledges the characteristics of smartphones, such as screen size and interaction method, which reflect the interface features of apps in detail.

2.2 Integrated-based frameworks

Insufficient literature on mobile phone characteristics concerning its interface feature has inspired the effort of an in-depth comprehension of the mobile interface feature. As a result, the abstraction levels for this type of framework are realized as an organization of mobile interface features.

In addressing the comprehensive aspect of usability issues on a mobile phone, Mugisha et al. articulate their framework in view of mobile phone UI practitioners [13]. Based on a review of usability principles, they defined five categories of UIs tailored for a feature phone. A pairwise comparison approach was used in mapping the UIs to usability principles. A checklist relevant to the UIs was developed to match the usability principles.

As a continuation, Xu and Jonsson [14] devised their framework by determining common interface features for tablet applications. The identified interface features were grouped into three categories: UI input, UI components, and UI characteristics. Each UI, which was paired with a developed checklist, was mapped to the usability principles based on their effect and relationship. Though tailored for tablet applications, and acknowledging smartphone characteristics in their work, the UIs mainly reflect desktop and web application UIs and features such as input, hardware, bookmarks, and headers.

2.3 Theoretical-based frameworks

The usability framework, which was developed based on usability conceptions of principles and criteria, mainly revolves around the effectiveness of existing usability measurements for evaluating apps. For example, Dubey and Rana [9] acknowledged the characteristics and features of mobile devices. They doubt the effectiveness of existing usability measurements on mobile phones. By hierarchically organizing usability indicators (principles), criteria, and properties based on a goal-mean relationship between the parameters, they formulated a framework for usability specialists to conduct an analytical evaluation of mobile phones. While focusing on the parameters of each abstraction level and all three categories of UIs (PUI, LUI, and GUI), their checklist suffers from redundancy, ambiguity, confusion, and indirectly measurable issues.

Pursuing a different approach, Gómez, Caballero, and Sevillano formed their framework by formulating a structure of heuristics and sub-heuristics, paired with a checklist based on their semantic relations [6]. They achieved excellent results in addressing mobile-specific usability issues while focusing on LUI and GUI. Unfortunately, though they argue for the effectiveness of a desktop-centered checklist for evaluating apps, a portion of their checklist stems from a web-based checklist that appears irrelevant for apps.

Judging by the limitations of mobile devices, Fatih Nayebi developed a heuristic-based framework for app evaluation [7]. A set of usability criteria established based on his review of academic and industrial heuristics, theories, and guidelines were assigned to the most relevant logical groups of the reviewed bibliographic references. Although he managed to address the characteristics of mobile devices, the proposed criteria were ambiguous and hardly addressed mobile interface features.

Further, arguing for the effectiveness of current usability measurement for mobile applications, Hoehle, Aljafari, and Venkatesh proposed a set of measurements for mobile applications in view of interface features based on measurement theory [12]. A content analysis approach was used to relate the constructs and variables. Though their work explicitly focused on apps, the measurements were tailored for Microsoft-based apps, and mobile interface features are not well addressed in their work. Instead, they emphasize aspects such as usability principles, aesthetics, and navigation.

2.4 Decision-based framework

The primary purpose of adopting a decision-based framework is to determine a usable mobile application. Lachgar and Abdelmounaim pursued an analytic hierarchy process in developing their framework. Grounded in measurement theory, he developed usability constructs and variables to facilitate the selection of usable mobile phones [15]. **Table 1** summarizes the literature review.

Earlier mobile usability studies emphasized the physical user interface. While the logical user interface persists across most computing platforms, rapid updates in smartphone technologies highlight the importance of the graphical user interface, particularly interface features. The coverage of IU studied previously conforms to the scope of UI covered in the reviewed framework from the age of feature phones to handhelds until smartphones, where PUI is scarcely studied in recent works.

3. Formulation of framework measurement

Representative definitions of usability by the industry (i.e., ISO 9241-11 [16], ISO 9126 [17]) and academia [5,18-22] are usually referred to most studies. In the context of mobile usability studies, Harrison et al. [5] work, which extends Nielsen's usability conception in view of the ISO 9241-11 context, is deemed as a comprehensive reference [23]. However, neither metrics nor checklists are associated with their work, thus leaving little support for usability

Table 1. Literature review summary.

Types of Framework	**Attribute-based frameworks**		**Integrated-based frameworks**		**Theoretical-based frameworks**				**Decision-based framework**
Authors	Hussain and Kutar [3]	Saleh, Ismail, and Fabil [8]	Mugisha et al. [13]	Xu and Jonsson [14]	Dubey et al. [9]	Gómez, Caballero and Sevillano [6]	Fatih Nayebi [7]	Hoehle, Aljafari and Venkatesh [12]	Lachgar and Abdelmounaim [15]
Viewpoint	developer	developer	developer	developer	Usability specialist	Non-expert	Usability specialist	Usability specialist	Non-expert
Aims of research	Conceptualise usability dimension	Conceptualise usability dimension	Bridging different groups of usability constructs	Bridging different groups of usability constructs	Bridging different groups of usability constructs	Addressing mobile usability issues	Conceptualise usability dimension	Bridging different groups of usability constructs	Select best alternatives among available usability criteria
Base structure	ISO 9241-11	PACMAD	Usability principles	Usability principles	Usability principles	Mobile constraints	Usability principles	Usability principles	ISO 9421-11
Mapping of abstraction levels components	Analytic Hierarchy Process	Goal Question Metric	Goal Question Metric	Pairwise-comparison	Pairwise-comparison	Content analysis	Content analysis	Content analysis	Content analysis
Context of use	Understanding measurement	Understanding measurement	Prioritizing constructs	Prioritizing constructs	Correlating constructs	Correlating constructs	Correlating constructs	Correlating constructs	Decision making
Benefit	Explicit measurement interpretation	Explicit measurement interpretation	Consistent judgement	Consistent judgement	Thorough construct classification	Thorough construct classification	Thorough construct classification	Thorough construct classification	Consistent judgement
Drawback	Tunnel vision bias	Tunnel vision bias	Large number of evaluations	Large number of evaluations	Tunnel vision bias	Tunnel vision bias	Tunnel vision bias	Tunnel vision bias	Large number of evaluations
Countermeasure implemented	Not applicable	Not applicable	Establishing selection criteria	Establishing selection criteria	Expert review	Not applicable	Not applicable	Expert review	Content rating
Platform	smartphone	smartphone	Feature phone	Tablet	Feature phone	Smartphone, tablet	smartphone	smartphone	Handhelds
Scope	LUI	LUI	LUI, GUI	LUI, GUI	PUI, LUI, GUI	LUI, GUI	LUI, GUI	GUI	PUI, LUI, GUI

inspection.

Consequently, this study formulates a framework for usability conception by reviewing checklist-related bibliographic references. This approach has been demonstrated in previous studies by conducting a bibliographic search when constructing their checklist-based framework [6,9,24]. They restricted their search scope by covering only the relevant and most influential references. In contrast, this study exhaustively examined relevant bibliographic sources for possible quality criteria denoting usability, such as standards, guidelines, and requirements in their work. This study reviewed requirements up to the evaluation life cycle to obtain a comprehensive description of apps' usability [25]. This process, however, is restricted to mobile and app-related sources.

Eleven relevant bibliographic references were reviewed for possible usability criteria. As a result, a collection of 572 measures was compiled. Measures irrelevant in the context of app usability were excluded to ensure mobile-specific measurements. **Table 2** highlights the distribution of redacted measures.

Measures referring to desktop-based input devices such as mouse and keyboard; web-related user interfaces such as a link to related content, breadcrumb, and splash screen; physical user interfaces such as a widget, soft keys, and notification drawer; and shared devices concern; performance-based checklists such as task completion time, loading time, download speed, and installation are removed from the collection. In addition, cross-domain concerns such as user experience and interaction design are excluded from the collection of candidate checklists. Technical and design aspects, such as naming convention and image size, which require coding inspection, are also removed. Any game-specific measure was removed due to the exceptional design objective, which distinguishes them from general apps such as banking, utilities, etc. [26,27].

Measures specifically for the impaired user, such as blind users, are also removed due to their exceptional design concerns. Conflicting measures within the same bibliographic reference were both excluded due to no concrete design decision. Measures referring to application purpose, e.g., "Application's pur-

Table 2. Distribution of redacted measures.

Exclusion criteria	**No. of excluded measures**
Design-related measures (Ex: image size)	34
Web-specific design elements (Ex: refresh button, wish list)	15
User-impaired related measures (Ex: visually impaired)	13
Game-based measures	10
Physical user interface related measure (Ex: widget, soft keys, notification drawer)	11
Programming related measures (Ex: naming convention)	9
Performance-based measures (Ex: time taken, number of successful task)	9
Input and output devices (Ex: Desktop based input hardware, wearable)	8
Miscellaneous (Ex: conflicting measure between the same bibliographic reference, application purpose, design statement or fact)	8
Cross-domain concern (Ex: user experience, interaction design)	4
Technical-related measures (Ex: installation, system resource)	3
Device specific features (Ex: shared device, tablet specific)	3
No. of redacted measures	127

pose is understandable at first sight", were excluded, although they refer to usability criteria of understandability. The rationale is that the measure does not contribute towards achieving user goals in operating an app, thus irrelevant in representing usability for apps.

Additionally, given that the measures consist of different forms such as usability requirements, heuristics, checklists, guidelines, recommendations, and usability problems, it is not possible to review the bibliography in terms of quality criteria that share a similar meaning, the same name, or both. Instead, regardless of their original form, the measures were rephrased into a checklist.

These measures were reviewed using a content analysis technique to develop usability constructs for apps. Content analysis of the measures developed relevant emergent quality attributes and interface features, which later resulted in a paired usability checklist. Initially, a conceptual definition for each usability criterion and interface feature is established. The conceptual definitions are made as unambiguous as possible in the context of apps. Conceptually similar items and repeated items referring to the same usability criteria are grouped together and rephrased to homogenize the resulting usability checklist. In the case of conflicting items, items that coincide with other items are retained, and the conflicting items are excluded from the checklist pool. Finally, the usability criteria are examined for similarities and differences in terms of their design patterns. The usability criteria are then grouped under conceptual units of similar apps' design patterns and usability features.

4. An integrated usability evaluation framework

Characterizing usability solely on usability principles or usability attributes suffers from a lack of reflection on interface features in detail such as notification and interaction method, which is another aspect influencing mobile usability. On the other hand, depending solely on the UI component for the evaluation would be inappropriate for measuring the usability factor. In addition, considering apps' short time-to-market, where usability specialists are rarely involved during the usability evaluation, there is a need to support non-usability specialists in conducting reliable usability evaluations from their point of view. These suggest a mobile usability framework that integrates multiple evaluator viewpoints. However, this would result in different evaluation criteria, such as interface features and usability features, in contrast to usability specialists and developers, who mostly view usability in terms of usability heuristics and quality criteria. **Figure 1** illustrates the conceptual framework.

The usability constructs are abstracted into three tiers of abstraction levels: usability feature level, usability criteria level, and interface feature level. Each abstraction level of the framework denotes a construct that consists of a group of framework components. The framework components are paired with the usability checklists for usability inspection.

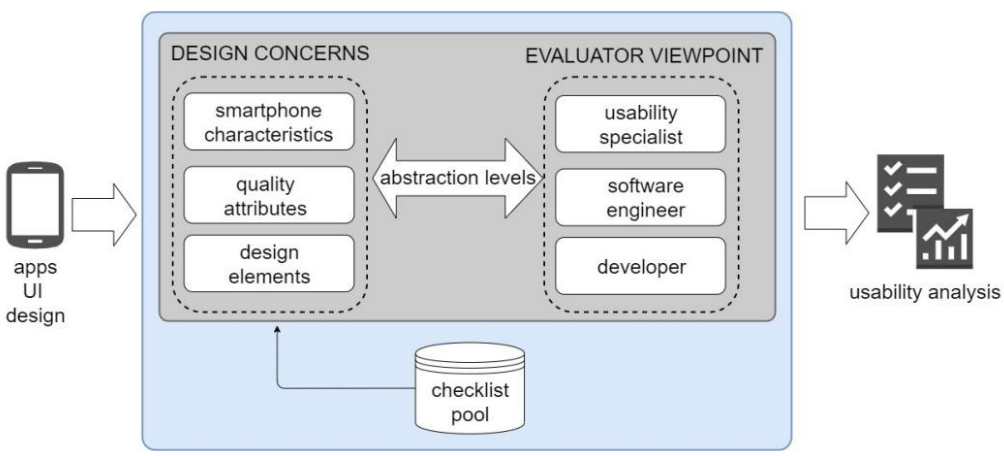

Figure 1. The conceptual framework.

Each framework tier reflects the different viewpoints of the usability evaluator and their level of expertise. The identified interface feature serves as the framework measurement, which formed the interface feature component for the lowest abstraction levels in the framework. The usability criteria are tied to the middle tier of the framework, the usability criteria component. Components for the top abstraction levels and usability features were identified by formulating conceptual units with similar usability criteria. **Figure 2** illustrates the framework abstraction level.

4.1 Usability features

Usability is commonly viewed by specialists in terms of constructs such as heuristics, principles, and guidelines, which are generally abstract. However, the mobile context of use, such as the interaction and operating environment, of the application on the intended platform has been regarded as an emergent property that affects usability [9,10,26]. Functional features of technology have been addressed in usability studies through design patterns [28-31]. The design pattern of app functionalities demonstrates the interaction complexities of smartphone apps. In this study, the design patterns are formulated as usability features to meet the viewpoint of usability specialists in conceptualizing usability as an emergent property of app interaction complexities. **Table 3** presents the elicited usability features in this study.

The usability features level denotes a collection of smartphone characteristics. These features are characterized by the attributes in the usability criteria level. It is formulated to meet the usability specialist's viewpoint in conceptualizing usability as an emergent property of app interaction complexities. It serves as an evaluation basis for both 1) specialists who view usability in terms of design patterns and 2) non-usability specialists, such as developers and designers, who could benefit from understanding usability in terms of design functionalities, in conducting usability evaluation.

4.2 Usability criteria

Characterizing usability solely by either usability principles or usability attributes suffers from a lack of reflection on interface features in detail such as notification and interaction methods, which is another aspect influencing mobile usability. However, on

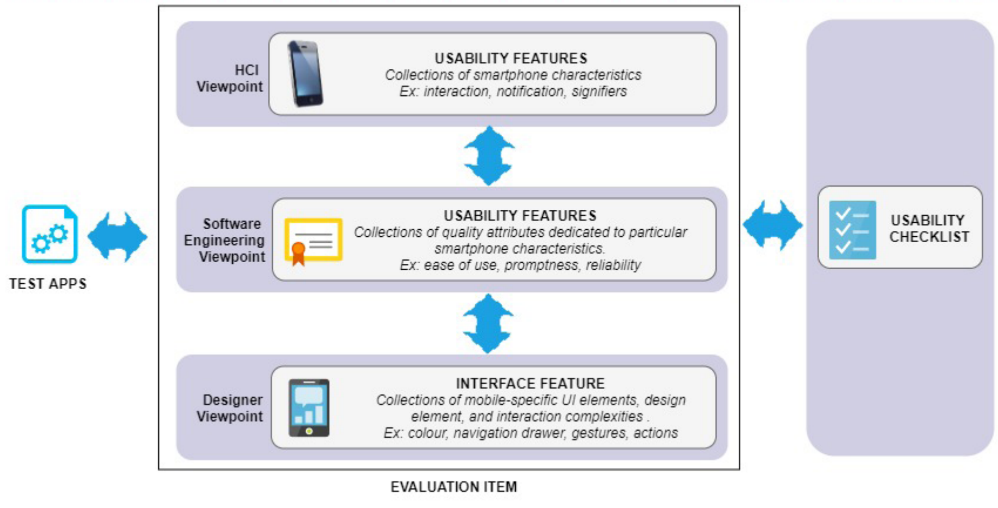

Figure 2. Framework abstraction level.

Table 3. Components of the usability features.

• F01 Interaction	• F04 Signifiers	• F07 Navigation	• F10 Data Entry
• F02 Notification	• F05 Aesthetic	• F08 Information Architecture	• F11 Workflow
• F03 Permission	• F06 Presentation	• F09 Search	• F12 Selection

the other hand, depending solely on the UI component for the evaluation would be inappropriate for measuring the usability factor. Therefore, this study bridged usability constructs, usability criteria, and interface features together with usability features.

The usability criteria level consists of a collection of usability attributes addressing the corresponding usability feature in the top tier. It emphasizes usability evaluation from a software engineering perspective. **Table 4** lists the components of the usability criteria.

The label next to each usability criterion denotes the usability features they are associated with. The usability criteria and interface features in the next tier facilitate usability evaluation and the perception of the evaluator in the domain of software engineering and development.

4.3 Interface features

The interface feature level defines components that are tied to the usability criteria in the middle tier. This level facilitates technical evaluators (e.g., analysts, designers, etc.) who perceive usability in view of the design context approach. It evaluates usability in view of design elements. Table 5 lists the components of the interface features.

In the formulated framework, each usability feature is decomposed into several usability criteria (as in a one-to-many relationship). However, a usability criterion is tied to more than one checklist, assessing different UI elements. Likewise, it is also possible for the UI elements to be associated with more than one usability criteria (as in a many-to-many relationship). **Table 6** exhibits the partial list of the paired usability checklist.

Table 4. Components of the usability criteria.

● Responsiveness (F01)	● Connectivity (F03)	● Readability (F06)	● Conciseness (F08)
● Interactivity (F01)	● Flexibility (F03)	● Relevance (F06)	● Structuredness (F08)
● Playability (F01)	● Security (F03)	● Accessibility (F06)	● Formality (F08)
● Ease of Use (F01)	● Visibility (F04)	● Trustworthy (F06)	● Effectiveness (F09)
● Safety (F01)	● Discoverability (F04)	● Navigability (F07)	● Accuracy (F10)
● Completeness (F02)	● Consistency (F05)	● Complexity (F07)	● Customisation (F10)
● Promptness (F02)	● Appropriateness (F05)	● Linkage (F07)	● Operability (F11)
● Reliability (F03)	● Familiarity (F05)	● Understandability (F08)	● Efficiency (F12)

Table 5. Components of the interface features.

● Action	● Button	● Icons	● Spinner
● Content	● Color	● Layout	● Snackbars
● Menus	● Default	● List	● Switches
● Layout	● Dialogue	● Navigation drawer	● System bar
● Steppers	● Expansion panels	● Picker	● Tabs
● Media	● Gestures	● Progress bar	● Text fields
● Action bar	● Grid list	● Slider sub-headers	● Typography
● Activity bar / circle	● Indicator	● Sub-screen	

Table 6. Partial list of the paired usability checklist.

Usability features	Usability criteria	Interface features	Checklist items
Navigation	Navigability	Action bar	Tabs or spinner in the top bar is used for quick view change
Navigation	Linkage	Action bar	Shortcut to most frequent task is provided
Notification	Completeness	Dialogue	There are at most 3 possible actions in a notification
Notification	Promptness	Dialogue	Notification is not created if it is possible for the app to recover from the error without user action
Presentation	Accessibility	Content	Content is structurally separated from navigational elements
Presentation	Relevance	Action bar	Unavailable action in the current context is hidden instead of disabled
Permission	Flexibility	Snackbars	The app allows to revert accidental activation
Permission	Security	Content	User's data are kept private and safe (encrypted in the event of loss or malfunction)
Signifiers	Visibility	Button	The UI Buttons are visible
Signifiers	Discoverability	Gestures	The user interface gives visual clues if something can be used with Pinch-To-Zoom gesture

5. Evaluating the frameworks usefulness

In our previous work, we validated the comprehensiveness of the framework components among academicians in Malaysia's public universities [32]. The components were refined based on the survey responses. Subsequently, the components were evaluated for their feasibility in real practice among software engineering practitioners in Malaysia and refined once again based on the survey response [33]. In this paper, we conducted an expert review and a semi-structured interview to evaluate the framework's usefulness in comparison to existing usability evaluation frameworks.

Usefulness is characterized in most usability studies as a composition of usability and as is utility [34,35]. Likewise, available usefulness questionnaires (e.g., USE and TAM) measure usefulness in the same dimension. The dimension includes a composition of several usability criteria, such as ease of use, learnability, and satisfaction, in addition to as-is utility. This section demonstrates the framework's evaluation in terms of its usefulness in comparison to the selected study.

The usability evaluation framework to be compared to the one from this study was selected through an exhaustive search of existing work on online databases subscribed by Universiti Putra Malaysia (UPM) and accessed publications. The search was performed using Google Scholar to review the recently proposed checklist-based framework published during the development of the framework in this study. The query returned 424 results in the English language. Any matching results that have been adopted in developing the framework in this study are omitted to avoid bias. Subsequently, publication on the checklist-based framework was filtered for selection. The process ends with two relevant search results. Since the work of Joseph [36] is more about usability heuristics, we have selected the work of Thitichaimongkhol and Senivongse [37] as a comparison against the formulated framework.

Methods and material

Prior to the evaluation, the participants were given a demographic form to record their background experience, the specifications of the smartphone used during the evaluation, such as brand and operating system, and their experience using apps in the dominant category in the marketplace.

Evaluating the entire framework measurement (373 checklists) from this study in comparison with the previous work is inefficient in terms of time and resources. Therefore, the evaluation scope covers usability measurements from both sets that match the

ISO/IEC 25010 product quality model. Although the usability criteria corresponding to both checklist sets for the evaluation have a different name compared to the ISO/IEC 25010 quality criteria, the conceptual definition for the corresponding usability criteria shares the same description as the ISO/IEC 25010 usability criteria.

Three apps from different categories commonly used by Malaysians (from survey responses in our previous study) are selected from the Play Store. Task analysis is performed on the apps to identify the primary task and the interface feature associated with each task. Usability criteria from both studies (this study and the other) corresponding to the interface features associated with the primary task are selected for the heuristic walkthrough. **Figure 3** exhibits an excerpt from the identified checklist from this study, corresponding to the usability criteria from ISO/IEC 25010 and interface features of the primary task for Lazada apps.

Likewise, the checklist for Set 2 is prepared using the same task and interface features of the same apps, corresponding to the same usability criteria as in Set 1. **Figure 4** exhibits an excerpt from the identified checklist from another study, corresponding to the usability criteria from ISO/IEC 25010 and interface features of the primary task for Lazada apps.

The participants were given two sets of checklists (76 items from the formulated framework and 39 items from the other framework), which correspond to the ISO/IEC 25010 usability criteria and interface features of primary tasks from selected apps. The evaluators were required to perform a heuristic walkthrough on three apps (Google+, Viber, and Lazada) using both checklist sets. Subsequently, they are required to review both checklist sets. The checklist sets are given in random order. The first evaluator is given Set 1, followed by Set 2. Meanwhile, the next evaluator is given Set 2, followed by Set 1.

Finally, both frameworks were rated for their usefulness using the USE questionnaires. The evaluators were given two sets of USE questionnaires, one for each framework. The questionnaire includes 30 checklist items on a 7-point Likert scale. The scale ranges from 1 (strongly disagree) to 7 (strongly agree). The resulting USE score was analyzed using paired t-test to determine if there was any difference between the compared frameworks. A post hoc test, Cohen's D, is used to investigate the effect size on the significance of the compared framework. Equation (1) explains Cohen's D measure of effect size.

Apps Usability Inspection Checklist - Lazada [Set 1]

Task 1: Search for an item

1) Look for a cat feeder.
2) Evaluate the interface involved during the process using the following checklist

Evaluation Item: 1) Gesture
(Interaction using hand motion. Ex: Swipe, tap, pinch)

CHECKLST	DESIGN COMPLIANCE			CHECKLIST RELEVANCE	
	Y	N	N/A	Y	N
1) Data entry does not require the use of both hand					
2) The user interface take into consideration of 2 handed thumbs usage					
3) The user interface take into consideration of right handed usage					
4) A virtual keypad is provided					
5) There is visual clues if something can be used with swipe-to-navigate					
6) The user interface can be controlled effortlessly with swipe-to-navigate					
7) It is possible for user to hit anywhere in a row to select target in tabular view					
8) The UI components size are big enough to allow users to effortlessly interact with their fingers					
9) Target is not placed too close to each other that user can easily hit the wrong one					
10) Voice assistance is provided					

Figure 3. Checklist for Set 1.

Apps Usability Inspection Checklist - Lazada [Set 2]

Task 1: Search for an item

1) Look for a cat feeder.
2) Evaluate the interface involved during the process using the following checklist

Evaluation Item: 1) Gesture

(Interaction using hand motion. Ex: Swipe, tap, pinch)

CHECKLST	DESIGN COMPLIANCE			CHECKLIST RELEVANCE	
	Y	N	N/A	Y	N
1) Are touchable objects (e.g., buttons) in the screen placed too close?					
2) Do objects on the screen have the size that is easy to touch (about 1 x 1 centimeter or 48 x 48 density independent pixels)?					
3) Does the system utilize screen space appropriately when displaying information by not using too much or too little padding or margin between elements.					
4) Does the system provide speech-to-text to support searching?					

Figure 4. Checklist for Set 2.

$$\delta = \frac{\overline{X}_2 - \overline{X}_1}{\sqrt{\frac{S_1^{\,2} + S_2^{\,2}}{2}}} \qquad (1)$$

where, $\overline{X}$ = sample mean, S = sample standard deviation.

Effect size is categorized into small, medium, large, and very large impacts [38]. An effect size of 0.2 indicates a small magnitude of the effect. Medium effect size ranges from 0.5 vicinities. Large effect size ranges from 0.8. Meanwhile, a very large effect size is indicated by values larger than, or equal to 1.3.

A semi-structured interview is conducted after the experiment to clarify why the experts rated one framework better than the other. The identity of each treatment, which framework was the one from this study, and which framework was from the previous study were not revealed until the end. The rationale is to have an expert's honest opinion on the framework.

6. Results and discussions

We approached eleven industrial experts, ranging from mobile developers to mobile testers and UX designers. However, five of them repeatedly rescheduled the dateline to complete the evaluation and failed to complete the requested evaluation even after more than three follow-up reminders.

Only six of the experts managed to complete the experiment. A difference in the overall USE scores rated by the six experts for both frameworks was computed. However, one of them gives an unreliable rating, even after reviewing the score given. It is not feasible to set up an upfront meeting with that expert. The expert's background profile showed that this expert is the only participant to select gaming apps as the most frequently used apps. Since gaming apps have different designs and purposes compared to other categories of apps, the expert's perception of usability might skew away from the other five participants, who were not familiar with mobile gaming. Thus, it is reasonable that the expert gave a contradictory score compared to the other participants. Therefore, the response by this expert was excluded from the analysis. The overall USE score collected from each expert is analyzed to evaluate the usefulness of the frameworks. **Table 7** exhibits the mean differences in the overall USE score for both treatments.

A paired t-test is used to determine the significance of the usefulness score with a p-value of 0.05. The distribution of the USE score differences between both groups of experts is normally distributed, thus making it appropriate for conducting a paired t-test. The mean indicates that the five experts (N = 5) gave a larger USE score for the formulated framework (mean = 180.60) compared to the other framework (mean = 156.60). In addition, a smaller standard deviation compared to the other framework indicates that the USE scores among the experts were more consistent in the formulated framework. **Table 8** exhibits the results of the paired t-test.

USE scores for the formulated framework are

Table 7. Overall USE score for both frameworks.

Paired Samples Statistics					
		Mean	N	Std. deviation	Std. error mean
Pair 1	Formulated framework USE score	180.60	5	10.991	4.915
	Previous framework USE score	156.60	5	15.143	6.772

Table 8. Results of the paired t-test.

Paired samples test							
		Paired differences			t	df	Sig. (1-tailed)
		Mean	Std. deviation	Std. error mean			
Pair 1	Formulated framework USE score – Previous framework USE score	24.000	20.273	9.066	2.647	4	.0286

24 points higher (mean paired difference = 24) than USE scores for the previous framework. There is enough evidence to claim that the mean USE score given by the experts for the formulated framework is greater than the previous framework, t (4) = 2.647, p = 0.0286. Thus, the null hypothesis is rejected since the p-value is less than 0.05. **Figure 5** illustrates the position of the calculated t statistic (within the H_0 rejection region), t-value, and p-value in a graph.

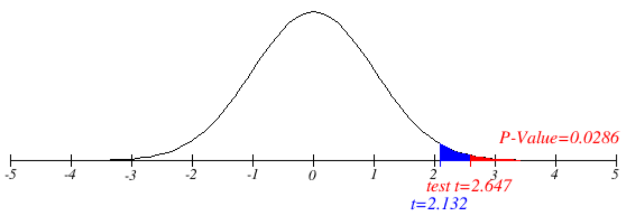

Figure 5. Position of calculated t statistics.

The learning experience gained in conducting the heuristic walkthrough using the checklist facilitates comprehending the framework measurement, thus ensuring reliable scoring of the framework's usefulness. However, although the result presented indicates that the USE scores of both frameworks are unlikely to occur by chance, the magnitude of the effect of the treatment (the formulated framework) over the other framework is unexplained. A post hoc test, Cohen's D, is used to investigate the effect size on the significance of the compared framework. The result of 1.81 indicates a very large effect size, implying a meaningful difference in the USE score between both frameworks.

Discussions

The demographic distribution of the experts indicates that they were the appropriate participants to evaluate the formulated framework measurement from the perspective of developers. Only two of the experts are experienced usability practitioners with seven years or more of experience in the field. Both of them were of different genders and were using different mobile OS. Additionally, all of the experts are of different ages, ranging from the twenties to the thirties and forties.

The formulated framework proved to be more useful than the previously proposed framework. The semi-structured interview revealed that the experts came to a consensus, agreeing that the formulated framework is more useful for usability evaluation compared to the other framework. The reason lies in the fact that the formulated framework measurement is much simpler, UI-oriented, and less ambiguous for experts, both developers and usability practitioners.

Both frameworks were compared using the same baseline measurement: the ISO/IEC 25010 product quality model (usability component) in conjunction with a common interface feature for the primary task in the evaluated apps. Nevertheless, the learning gained from experiencing the framework measurement exhibits the usefulness of each framework.

The formulated framework is criticized for its large number of checklists. However, it is not practical to inspect every available criterion. Usability evaluation is commonly conducted based on an evaluation plan established beforehand, which determines the criteria to be evaluated during an inspection. The formulated framework came in handy due to its features in supporting different backgrounds of usability evaluators through the abstraction level. In fact, restricting usability evaluation to usability criteria of interest will eventually reduce the number of checklists to be used during the usability evaluation.

7. Threats to validity

Threats are inevitable yet manageable in research. In this section, threats to internal, external, conclusion, and construct validity are discussed. The selection of associated usability criteria tied to UI elements was determined by adopting the ISO/IEC 25010 product quality model (usability component) as a benchmark criterion in comparing the formulated framework over previously proposed frameworks. However, the experiment is still vulnerable to the order effect. Thus, in replicating the experiment over the other framework in comparison, two sets of evaluation plans representing the formulated framework and the previously proposed framework were given to the evaluator in random order.

Regarding the external threat, the respondent's expertise and experience in using the evaluated app might affect the validity of the result. The respondents consist of field experts from various branches of software engineering disciplines and app development stages with a different range of years of experience. In addition, they might use their experience of using a particular type of app, e.g., transactional, communication, or games, as a benchmark in scoring UI elements. Altogether, the respondent might perceive usability differently based on their background of expertise and experience with the app, thus affecting their subjective judgement. These threats are controlled through two countermeasures. Firstly, a conceptual definition of the evaluated interface feature was established. There is a possibility that an interface feature is recognized by a different name in academia and industry. For example, a drop-down menu is well-known in desktop computing. On some occasions, it is used as a jump menu. However, in mobile computing, the drop-down menu is recognized as a list. In addition, the jump menu is known as a spinner in mobile computing. Furthermore, elements such as sub-screen and gesture are interface features that are absent in desktop computing and could be misinterpreted differently by individuals. This necessitates a further description of a UI element's operation or behavior in view of desktop computing to facilitate an inexperienced evaluator in this case. Secondly, the experiment is designed as a repeated measure to reduce variability across participants.

The main threat to the conclusion validity of the result is statistical power. This threat is alleviated by applying the most common statistical test, appropriate for the research design of within-subject design. Moreover, the significance level was 5%. Hence, the chance of a Type I error is small.

A checklist from the previous study is used in comparison to the checklist in this study to manage construct validity. The scores of both checklists in measuring the ISO/IEC 25010 product quality model (usability component) were correlated in conjunction with the use of an established questionnaire to measure the framework's usefulness. In addition, a well-established usability questionnaire was carefully selected for this study to measure usefulness appropriately.

8. Conclusions

This study empirically evaluates the usefulness of an integrated usability evaluation framework through an expert review. The framework measurement is reviewed and compared against a framework from another study. Both frameworks were compared based on the ISO/EIC 25010 product quality model (usability component). Hypothesis testing was conducted to investigate the significance and effect size of the response from the expert review. The results of the statistical test proved that the formulated framework had a significant and large effect size and was more useful compared to the other framework. In the future, we plan to improve the effectiveness of this framework by comparing the results of using it in usability testing against usability inspection. The rationale is to alleviate a possible false alarm in the formulated framework measurement and capture the true usability problem. Consequently, an additional checklist could be proposed based on the usability testing result to complement the developed usability measurement.

Author Contributions

H. R. conceived the idea and study of proposing a usability evaluation framework for mobile apps that incorporates the usability criteria and interface features in conjunction with different evaluator viewpoints into a framework abstraction level. H. Z., A. K. and the late A. A. A. A. served as H.R.'s supervisor and co-supervisors on her Ph.D. thesis at the Universiti Putra Malaysia. All authors reviewed and approved the final manuscript.

Conflict of Interest

There is no conflict of interest.

Funding

This research was partially funded by the Research University Grant Scheme (RUGS), Universiti Putra Malaysia (UPM).

References

[1] Moumane, K., Idri, A., Abran, A., 2016. Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards. SpringerPlus. 5, 1-15.

[2] Hussain, A., Abubakar, H.I., Hashim, N.B. (editors), 2014. Evaluating mobile banking application: Usability dimensions and measurements. Proceedings of the 6th international Conference on Information Technology and Multimedia; 2014 Nov 18-20; Putrajaya, Malaysia. New York: IEEE. p. 136-140.

[3] Hussain, A., Kutar, M. (editors), 2009. Usability metric for mobile application. 2008 Proceedings of the 10th International Conference on Information Integration and Web-Based Applications and Services (IiWAS'08); 2008 Nov 24-26; Linz, Austria. New York: Association for Computing Machinery. p. 567-570.

[4] Liu, N., Yu, R., 2017. Identifying design feature factors critical to acceptance and usage behavior of smartphones. Computers in Human Behavior. 70, 131-142.

[5] Harrison, R., Flood, D., Duce, D., 2013. Usability of mobile applications: literature review and rationale for a new usability model. Journal of Interaction Science. 1, 1-16.

[6] Yáñez Gómez, R., Cascado Caballero, D., Sevillano, J.L., 2014. Heuristic evaluation on mobile interfaces: A new checklist. The Scientific World Journal. 1-19.

[7] Nayebi, F., 2015. iOS application user rating prediction using usability evaluation and machine learning [Ph.D. thesis]. Quebec: University of Quebec.

[8] Saleh, A., Ismail, R., Fabil, N. (editors), 2017. Evaluating usability for mobile application: A MAUEM approach. ICSEB 2017 Proceedings of the 2017 International Conference on Software and E-Business; 2017 Dec 28-30; Hong Kong. New York: Association for Computing Machinery. p. 71-77.

[9] Dubey, S.K., Gulati, A., Rana, A., 2012. Integrated model for software usability. International Journal on Computer Science and Engineering. 4(3), 429.

[10] Elsantil, Y., 2020. User perceptions of the security of mobile applications. International Journal of E-Services and Mobile Applications (IJESMA). 12(4), 24-41.
DOI: https://doi.org/10.4018/IJESMA.2020100102

[11] Malatini, S., Bogliolo, A. (editors), 2015. Gamification in mobile applications usability evaluation: A New Approach. MobileHCI'15: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services; 2015 Aug 24-27; Copenhagen, Denmark. New York: Association for Computing Machinery. p. 897-899.

[12] Hoehle, H., Aljafari, R., Venkatesh, V., 2016.

Leveraging Microsoft's mobile usability guidelines: Conceptualizing and developing scales for mobile application usability. International Journal of Human-Computer Studies. 89, 35-53.

[13] Mugisha, A., Nankabirwa, V., Tylleskär, T., et al., 2019. A usability design checklist for Mobile electronic data capturing forms: the validation process. BMC Medical Informatics and Decision Making. 19(1), 1-11.
DOI: https://doi.org/10.1186/S12911-018-0718-3

[14] Xu, H., Jonsson, M., 2012. Tablet application GUI usability checklist—Creation of a user interface usability checklist for tablet applications [Master's thesis]. Huddinge: Södertörns University College.

[15] Lachgar, M., Abdelmounaim, A., 2017. Decision framework for mobile development methods. International Journal of Advanced Computer Science and Applications. 8(2), 110-118.
DOI: https://doi.org/10.14569/IJACSA.2017.080215

[16] ISO 9241-11:1998 Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 11: Guidance on Usability [Internet]. International Organization for Standardization; 1998 [cited 2018 Dec 21]. Available from: https://www.iso.org/standard/16883.html

[17] ISO/IEC 9126-1:2001 Software Engineering—Product Quality—Part 1: Quality Model [Internet]. International Standard for Standardization; 2001 [cited 2018 Dec 21]. Available from: https://www.iso.org/standard/22749.html

[18] Nielsen, J., Budiu, R., 2012. Mobile usability. New Riders Press: Berkeley CA.

[19] Constantine, L.L., Lockwood, L.A., 1999. Software for use: A practical guide to the models and methods of usage-centered design. Addison-Wesley Publishing Co.: Boston.

[20] Böckle, M., Rühmkorf, J. (editors), 2019. Towards a framework for the classification of usability issues. Human-Computer Interaction-INTERACT 2019: 17th IFIP TC 13 International Conference; 2019 Sep 2-6; Paphos, Cyprus. p. 610-614.

[21] Kronbauer, A.H., Santos, C.A., Vieira, V. (editors), 2012. Smartphone applications usability evaluation: A hybrid model and its implementation. Human-Centered Software Engineering: 4th International Conference, HCSE 2012; 2012 Oct 29-31; Toulouse, France. p. 146-163.

[22] Olsina, L., Santos, L., Lew, P., 2014. Evaluating mobileapp usability: A holistic quality approach. Lecture notes in computer science. Springer, Cham.: New York. pp. 111-129.

[23] Fabil, N.B., Saleh, A., Isamil, R.B., 2015. Extension of pacmad model for usability evaluation metrics using goal question metrics (Gqm) approach. Journal of Theoretical and Applied Information Technology. 79(1), 90-100.

[24] Inostroza, R., Rusu, C., Roncagliolo, S., et al. (editors), 2012. Usability heuristics for touchscreen-based mobile devices. 2012 Ninth International Conference on Information Technology—New Generations; 2012 Apr 16-18; Las Vegas, NV, USA. New York: IEEE.

[25] Mi, N., Cavuoto, L.A., Benson, K., et al., 2014. A heuristic checklist for an accessible smartphone interface design. Universal Access in the Information Society. 13, 351-365.

[26] Soomro, S., Ahmad, W.F.W., Sulaiman, S. (editors), 2012. A preliminary study on heuristics for mobile games. 2012 International Conference on Computer and Information Science; 2012 Jun 12-14; Kuala Lumpur, Malaysia. New York: IEEE. p. 1030-1035.

[27] Zahra, F., Hussain, A., Mohd, H. (editors), 2017. Usability evaluation of mobile applications; Where do we stand? The 2nd International Conference on Applied Science and Technology 2017 (ICAST'17); 2017 Apr 3-5; Kedah, Malaysia.

[28] Zamfiroiu, A., 2014. Factors influencing the quality of mobile applications. Informatica Economica. 18(1), 131.

[29] Homann, M., Wittges, H., Krcmar, H., 2013. Towards user interface patterns for ERP applications on smartphones. Business Information Systems. 157, 14-25.

[30] Roder, H. (editor), 2012. Specifying usability features with patterns and templates. 2012 First

International Workshop on Usability and Accessibility Focused Requirements Engineering (UsARE); 2012 Jun 28; Zurich, Switzerland. New York: IEEE. p. 6-11.

[31] Punchoojit, L., Hongwarittorrn, N., 2017. Usability studies on mobile user interface design patterns: A systematic literature review. Advances in Human-Computer Interaction. 1-22. DOI: https://doi.org/10.1155/2017/6787504

[32] Rahmat, H., Zulzalil, H., Ghani, A.A.A., et al., 2018. A comprehensive usability model for evaluating smartphone apps. Advanced Science Letters. 24(3), 1633-1637.

[33] Zulzalil, H., Rahmat, H., Ghani, A.A.A., et al., 2019. Conceptualising mobile apps usability dimension: A feasibility assessment of Malaysian industrial practitioners. International Journal of Engineering and Advanced Technology. 9(1), 1708-1713.

[34] Tarkkanen, K., Harkke, V., Reijonen, P., 2015. Are we testing utility? Analysis of usability problem types. Design, user experience, and usability: Design discourse. Springer: Cham. pp. 269-280. DOI: https://doi.org/10.1007/978-3-319-20886-2_26

[35] MacDonald, C.M., Atwood, M.E. (editors), 2014. What does it mean for a system to be useful? Proceedings of the 2014 Conference on Designing Interactive Systems; 2014 Jun 21-25; Vancouver BC, Canada. New York: Association for Computing Machinery. p. 885-894.

[36] Joseph, V., 2017. User experience guidelines for improving retention rate in mobile apps [Master's thesis]. Madrid: Universidad Politécnica de Madrid.

[37] Thitichaimongkhol, K., Senivongse, T., 2016. Enhancing usability heuristics for android applications on mobile devices. Proceedings of the World Congress on Engineering and Computer Science. 1, 19-21.

[38] Sullivan, G.M., Feinn, R., 2012. Using effect size—or why the P value is not enough. Journal of Graduate Medical Education. 4(3), 279-282.