Usability Testing of Smartphone Applications: a Literature Review

1. Introduction

Smartphone applications have become part of people's lives and will be even more integrated in the near future. With smartphone's increasing popularity and incomparable ubiquity, many researchers from various fields adopt this platform to design and develop applications as a novel solution to their concerning issues.

Usability as the core of product design, including smartphone applications, has been broadly recognized in computer engineering and HCI fields (Nielsen,1993; Brooke,1996; Dumas, et al, 1999; Preece, 2000; Shneiderman, 2000). The definition of usability was probably first attempted by Miller (1971) in terms of measures for "ease of use", and these were developed further by Bennett (1979) to describe usability. The concept of usability was first fully discussed and a detailed formal definition was attempted by Shackel (1981), and Bennett (1984) modified and developed the definition. Nielsen (1994) defines usability as "the question of how well users use that functionality" and "how well" can be systematically approached, improved, and evaluated (possible measured) by five attributes: learnability, efficiency, memorability, errors, and satisfaction.

According to the latest ISO 9241-11(2018), usability, as one outcome of interaction, is "the extent to which a system, product, service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." This definition describes a concept that is more comprehensive than "ease-of-use" or "user-friendliness". In this updated definition, the scope has been extended to system, product, and service instead of just product. In this paper, "product" refers to "system, product, and service".

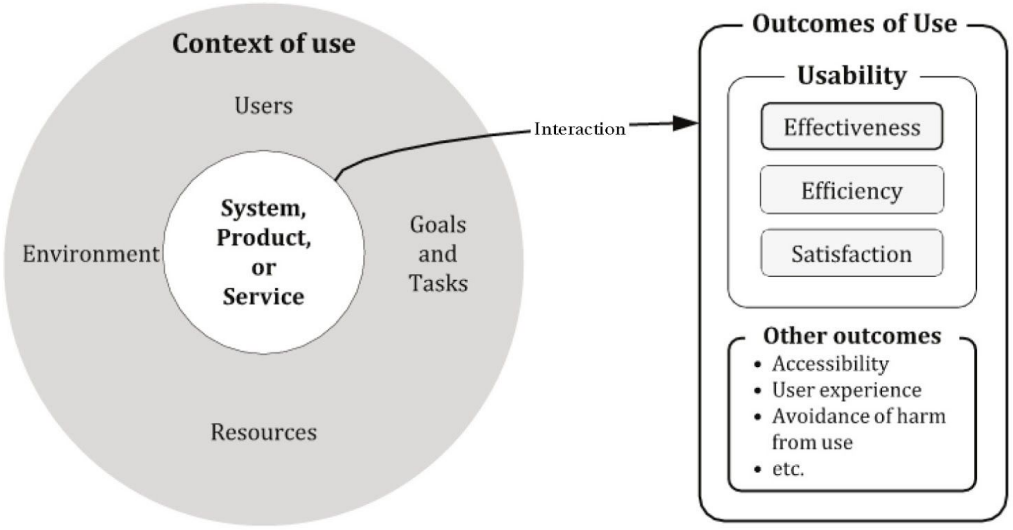


Figure 1: Usability and other outcomes of use, adopted from ISO 9241-11, 2018

As illustrated in Figure 1, a product represents the object of interest. The object of interest is shown as the center of the context of use, which consists of the users, the goals and tasks, the resources, and the environment. Tasks are a set of activities undertaken to achieve a specific goal. Usability shown as an outcome of the interaction is composed of effectiveness, efficiency, and satisfaction. Other outcomes of the interaction are accessibility, user experience and avoidance of harm from use. Here user experience refers to user's perceptions and responses

from the use of a product. The extent to which usability is achieved will vary depending on the characteristics of the product, the goals, the tasks, the users, the resources, and the user environment.

Usability testing, just as its name implies, is a test of the usability of a product with users. In the past, usability testing was conducted in labs by specialists with numerous subjects. The cost of such usability testing prohibited the technique to be widely used. This situation shifted in the early 1990s when researchers (e.g., Nielsen, 1994; Virzi, 1990 and 1992; Lewis, 1994) showed that smaller numbers of participants were sufficient to discover most of the usability issues. This finding has boosted employment and evolution of usability testing. Empirical usability testing is one of four methods (automatical, empirical, formal, and informal) for software usability inspection (Nielsen, 1994), and it has developed into a rich area of study with established procedures and ever-changing testing devices, thanks to rapid technological development.

There are two types of empirical usability testing, formative and summative. Formative usability testing focuses on problem discovery in an iterative design process and is commonly used to test product prototypes with a small number of users (Gould and Boies, 1983; Gould and Lewis, 1985; Gould et al., 1987; Gould, 1988;). Techniques for formative usability testing are elaborated in several handbooks (e.g., Dumas, Dumas, & Redish, 1999; Rubin, Chisnell, & Spool, 2008; Barnum, 2010). Summative usability testing is more formal and focuses on obtaining scientific measurements, mostly on a finished product, often with a goal to test certain hypotheses. Summative usability testing is widely adopted in academics in many fields, such as education (Navarro, Molina, & Redondo, 2015), healthcare (Kascak, et al; Reis et al., 2019); weather monitoring (Adhy, Prasetio, Noranita, & Saputra, 2018), geographic information system (Unrau & Kray, 2019), etc. Ideally, a project for designing a product with good usability would employ both types of usability testing in the iterative design and development process (Lewis, 2006).

This review will focus on empirical usability testing covering various types of products, especially on smartphone applications. This is because smartphones are expected to be the primary device for users to access information or other functionalities in the future. A handful of review articles have been published in recent years on mobile app usability testing (Alva et al., 2003; Zhang & Adipat, 2005; Coursaris & Kim, 2006; Alshamari & Mayhew, 2009; Bastien, 2010; Nayebi et al., 2012; Harrison et al., 2013). These works cover the entire methodology of usability testing, giving only limited attention to usability attributes and their measurements. Alturki and Gay (2019) reviews usability's attributes for mobile applications, but they only review eighteen papers. The purpose of the essay for the Comprehensive Exam is to address this shortcoming by symmetrically reviewing attribute measurements used in the studies of smartphone applications from 2017-2019, with plans to extend the review to earlier years if needed in the future. While the focus is on smartphone application usability attributes and measures (with a special focus on "satisfaction" and its measures), the review also identifies, analyzes, and summarizes other aspects of the methodologies employed by researchers, including research ethics and policy, experiment settings, sampling/recruiting, experiment methods, novel usability testing devices used in the tests. Another goal is to link theories in cognitive psychology to relevant usability attributes used/studied in the studies that are reviewed. This review covers a total of 40 publications with smartphone application usability testing published from 2017 to 2019.

2. Method of paper selection

Firstly, queries such as (HCI OR usability) AND mobile phone AND application AND (evaluation OR experiment OR study OR testing, (HCI OR usability) AND ethics were used to search for papers from 2017 to

2019 in three databases, IEEE, ACE DL, and Scopus. Titles and abstracts of the returned papers were manually examined to select publications for review.

Publications met the conditions below were chosen:

- Research that adopts the usability testing method to assess a mobile phone application(s)/prototype(s).
- Full research papers.
- Written in English

Publications of the following kinds were excluded:

- Papers that only present the design process and implementation of applications, but no usability testing.
- Papers that evaluate games.
- Papers that evaluate one feature(s) or pattern(s) that can be employed across different applications or mobile phone operating systems, iOS, Android.
- Papers that evaluate usages of new technology, for instance, voice-enabled technology, that can be applied to mobile phones.
- Papers that recommend design principles or techniques without empirical data support.
- Introductory level papers on usability testing methodologies or metrics.
- Papers that present usability testing tools only.

In the end, 40 papers were selected for this review.

3. Usability Related Concepts: Acceptability, Usefulness, Utility, User Experience

Nielsen in his highly influential (over 20k citations) work (1994) emphasizes that system acceptability has many components (Figure 2). Usability is only one of many and could be traded off with others under various circumstances. For example, on Youtube, commercials are often played before the videos but the user can skip it after 5 seconds. The 5-second skipping rule is a strategy to achieve a compromise between usability and other system acceptability components to reach an adequate level of acceptability.

Figure 2: System acceptability components, adopted from Nielsen, 1993.

Nielsen defines system acceptability as "the question of whether the system is good enough to satisfy all the needs and requirements of the users and other potential stakeholders, such as the user's clients and managers". And it can be further broken down into two categories: social acceptability and practical acceptability. Given a system has been accepted by society, we can further analyze its practical acceptability along with a number of dimensions: cost, support, reliability, compatibility with existing systems, and usefulness, etc.

Usefulness is "whether the system can be used to achieve some desired goals". It has two aspects: utility and usability. Utility asks "the question of whether the functionality of the system in principle can do what is needed ", and usability is about "the question of how well users use that functionality". In Nielsen's model, usability and utilities are two distinct components of acceptability. He touches on that usability is a concept highly related to the user in his conceptualization of system acceptability.

Figure 3: Usefulness, Utility, Usability, User Experience

Usability and user experience in ISO 9241-11(2018) are two of the four intended outcomes of interaction (Figure 1). Compared with usability, user experience is a newer term and was brought to wider recognition in the 1990s (Norman, Miller, & Henderson, 1995). "User experience" is "a person's perceptions and responses that result from the use and/or anticipated use of a system, product or service"(ISO 9241-11, 2018). Recall, ISO 9241-11(2018) expounds usability as "the extent to which a system, product, service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use". This suggests that in ISO 9241-11 conceptualization, usability and user experience are seen as different aspects associated with user interaction with a product. Usability focuses more on the *product's* effectiveness, efficiency, and satisfaction to users, while user experience covers more a *user's* subjective perceptions and responses resulted from the interaction with a product, and its context (e.g. brand name, services, etc.).

In my view, since user satisfaction is a core element in usability, it is meanwhile part of a person's perceptions and responses that result from the interaction with a system. Therefore it is an attribute that can help assess both usability and user experience. Based on this observation, usability and user experience cannot be two disjointed concepts, instead, they overlap at least on the attribute of satisfaction in the context of intended outcomes of interaction.

ISO definition of user experience is not the only definition of user experience. Nielsen and Norman defines user experience as "all aspects of the end-users interaction with the company, its services, and its products" (Nielsen & Norman, 2014). This view is accepted by many user experience researchers and practitioners (Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). In this more holistic view of user experience, usability is considered as one contributing factor to the overall user experience with a product. We agree more with Nielson and Norman's definition of user experience. The relationships among usefulness, utility, usability and user experience are illustrated in Figure 3.

4. Summative usability testing procedure suitable for academic research and publications

Usability testing is one part of the product design process. Normally, formative usability testing is addressed at the early or middle stage of product design, while summative usability testing is at the end of product design. During the research and prototyping phase, a number of design methodologies, theories, models, and principles can be applied, for example, user-centered design (Garrett, 2010), iterative design (Gossain, 1990; Nielsen, 1993), user experience design (Pine & Gilmore, 2011; Mendoza, 2013; Newbery & Farnham, 2013; Levin, 2014), emotional design (Norman, 2004), activity theory (Kaptelinin & Nardi, 2006), conceptual model (Johnson & Henderson, 2011), Norman's seven principal (Norman, 1988) and principles rooted from psychology (Johnson, 2013; Evans, 2018), design languages (Hoober & Berkman, 2011; Nielsen & Budiu, 2013; Neil, 2014), etc. In addition, a handful of research methods in the HCI field are on the plate to be selected,

e.g. focus-user group, card sorting, fast-prototype, etc (Farrell, 2017). Researchers can choose and combine them based on their research questions and restrictions. Once the application is designed and implemented to a high-fertility prototype or a fully functional product, summative usability testing can be initiated.

The summative usability testing method is rooted in the classic controlled experiment methodology with a hypothesis, random selecting, tight control, controlled groups, and a sufficient size sample. This classic approach is challenging in a fast-speed, high-pressure environment in the business world, therefore formative usability testing is more popular in design studios and consulting companies. While in an academic environment where pressuring profit is not the primary or the most urgent goal, a classic and summative approach has its space to be adopted to obtain quantitative proof and qualitative insights for research questions. In this section, a general procedure of summative usability testing suitable for academic studies and publications is presented, which can also be applied to mobile phone applications.

Step 1: Setting up the objective(s) of usability testing

The primary objective is to evaluate the usability of a mobile phone application(s) and make improvements before it is released (Dumas and Redish, 1999). The objective can be as a whole as one research question or be broken down into several research questions based on the research scope. For example, in the Measurement Record (MR) study I conducted under the supervision of Dr. Cui, one of the objectives is to asset its learnability by measuring whether the MR is intuitive enough to be used without instructions, while another is to evaluate its functionality by finding out whether certain features in MR help user's input to converge (Hong et al., 2020 (in preparation)).

Step 2: Defining usability attributes and measures based on the objective(s)

Based on the objective (s), attributes and measures of the attributes need to be defined. Usability attributes are the dimensions of usability, while measures are the metrics that measure the related attributes. A good number of questionnaires have been established that provide measures for usability attributes for generic usability testings. The attributes and measures, including questionnaires, will be elaborated in Section 4. To better interpret usability measurements, user demographics and relevant background information are often collected as well. For example, in the Add2Ontology usability testing, information on the user's prior experience using Wiki was collected, which helps to explain the user's preference for WikiData as a platform for building a domain ontology (Hong et al., 2020 (in preparation)).

Step 3: Designing the usability testing experiments

Experiment design involves a set of independent variables, dependent variables, and experiment participants. The four essential components of a usability testing experiment: user groups, task, the application(s), and the test settings, can be mapped to these elements: tasks are independent variables, the application (s) is an independent variable, a test setting is a dependent variable, while user groups are experiment participants.

Usability data can be collected using a variety of methods, including surveys/questionnaires, interviews, system activity log, direct observation, video recording, think-aloud sessions, retrospective testing, and constructive interaction coaching method, etc. Both quantitative (e.g., task completion time, number of errors) and qualitative data (e.g., agreement on the ease to use) are often collected.

Users groups

Users groups are a sample of intended users, whose profiles/characteristics should have already been created at the starting stage of product design. These characteristics can include the level of experience, personality traits (Jungian personality types, field dependence/independence, locus of control, imagery, spatial ability type A/B personality, ambiguity tolerance), demographic characteristics (sex, age), education level, etc (Aykin & Aykin, 1991). The level of experience, both with the application and in the domain of interests, is one of the most critically determining characteristics. In a usability testing of chemistry dictionary application (ChemDic) developed on Android studio, the users group has two following characteristics: 1) own an android device and 2) recently involved in chemistry learning (Nazar & Zulfadli, 2017). Usually, a screening questionnaire is designed to verify participants' characteristics. A number of participants will be recruited and be divided into groups. User sampling is covered in Step 5.

Task scenarios

The most important consideration of test tasks in a usability testing is to cover all types of tasks real users will perform with the application to achieve their goals. After listing all possible tasks, category them into core and peripheral. The researcher can create several test task scenario(s) for users to perform these tasks. Do ensure that all tasks are covered by the scenarios regardless of duplications of some tasks in different scenarios. Here is an example of a task scenario:

> *It's about Black Friday and you want to buy product A as a gift for Christmas. Please use the application to purchase product A, and ship to your home. <Address: ***; Phone number: ***; Credit Card: ***;>*

A task should not be described as a series of concrete operations for the user to carry out. For example, a test task for a calendar app could be creating an event and sharing it with friends, not " click the '+' button", " input a name for an event", etc. In a usability testing of a mHealth system, seven tasks were selected for covering the full range of the system's functions and varying levels of complexity (Hughes et al., 2019).

Application(s)

In usability testing, there can be more than one application or multiple versions of one application and can be viewed as an independent variable in experiment design. Nurhudatiana et al., (2018) studied the usability of two versions of an application, desktop, and mobile phone.

Testing setting

The testing setting should mimic the natural environment in which the user would normally be while using the application. The testing setting can be in the lab, or in the field. The optimum lab setting is a testing room and an observation room separated by a one-way mirror. The testing room is a simulation closed to the real usage environment with cameras and microphones for recording purposes. The observation room has equipment for researchers to observe the testing. Yet, a simple setting with a table, a chair, and a mobile phone can also be used for a usability testing lab as well, depending on the experiment. Conducting usability testing in the field is to allow users to employ an application(s) in their real environments, sometimes, for a period of time. The participants in a study of mobile support for older adults and their caregivers (Quinn, Staub, Barr, & Gruber-Baldini, 2019) used the application in their daily work and life for a one-month period. Usually, usability testing only chooses one setting for the entire setting, and the setting is considered as a control variable.

Experimental design

Now we have tasks, user groups, an application (s), a test setting, in usability testing. Experimental design methods, between groups, within groups, or matched pairs can be used to design the usability testing. Also, the order of tasks and versions of an application can be Latin square to counterbalance the order effect.

Step 4: Preparing
usability testing materials.

After the experiment is designed, all experiment materials, e.g., a recruitment letter/email/poster, an instruction sheet, an informed consent form, a screening questionnaire, a post-experiment questionnaire, etc. need to be prepared and an Institutional Review Board (IRB) approval is required. Pilot testing is necessary beforehand to discover any problems and make appropriate adjustments in experiment instruments/settings in a timely manner.

Step 5: Sampling/Acquiring participants

At this point, everything is ready besides recruiting users to the participant in the usability testing based on the user profiles mentioned in Step 3. In this step, we clarify how to acquire participants and the number of participants required for usability testing.

There are some common sources where researchers can advertise their recruitment information, application domain related agencies, market research firms, and design consulting firms, college campuses, email lists, personal/professional networks. The researcher should choose one or multiple sources based on the application(s). For instance, if a mobile phone application (Zaror et al., 2019) is designed for community-based surveillance of traumatic dental injuries, then local dental clinics are the primary resource for researchers to recruit participants. If an application is an application for the general public, e.g. an application that assists you to find a vacant parking space (Ng, Cheong, Hajimohammadhosseinmemar, & Yap, 2017), any individual who owns a car is a potential user, and randomly recruiting participants on parking lots could be an inexpensive way to get participants. After advertising the recruiting in each resource for potential participants, people who are interested in participation will get in touch with researchers. Researchers can distribute the screening questionnaire prepared in Step 4 to qualify and select participants.

There are two types of samples, probability samples, and nonprobability samples. A probability sample is a sample in which every unit in the population has a chance (*greater than zero*) of being selected in the sample, and this probability can be accurately determined. The methods to obtain a probability sample include simple random sampling, systematic sampling, stratified sampling, probability proportional to size sampling, cluster sampling, and multistage sampling, etc. A nonprobability sample is a sample where some elements of the population have *no* chance of selection. The methods include convenience sampling, quota sampling, and purposive sampling.

The number of participants for usability testing depends on many factors. Alreck and Settle (1985) proposed a summary of factors indicating the appropriate use of large and small samples.

- Decisions based on the results are major and may cause serious or costly consequences, OR decisions are minor and have few serious consequences.
- High confidence required in results OR rough estimates accepted
- The important measures have high variance OR low variance.
- Analyses will require dividing the entire sample into small groups, OR no require of small groups
- The cost and time of a usability test will not increase dramatically with sample size, OR it will.
- Time and resources are available; OR not.

In some cases, power analysis can be conducted to determine the sample size, but more often than not, a sample of 20-40 participants are often used, due to a lack of prior knowledge on expected variances in different attributes to be measured and the difficulty in recruiting a large number of participants in general.

Step 6: Test session

After the recruitment, all participants perform the experiment. The test session normally consists of an introduction, task performance, post-task activities, and sometimes debriefing. In the beginning, a brief introduction is given to participants. Mostly it contains the purpose of the test with an emphasis on its goal to test the product, not the participants, the right to drop out at any point without penalty, what data will be collected and the fact that all results will be confidential. Sometimes, the demonstration of the application and a brief of the procedure of the test is given and a period of time is provided for participants to get familiar with the mobile phone and application. Then an informed consent form and a detailed instruction sheet will be distributed to the participants. During the introduction, participants are free to ask any questions. After the introduction, all participants follow the instruction sheet to perform tasks on the application(s) and fill out questionnaires independently or collaboratively, with researchers observing or not. Normally, researchers provide no assistance during the phase unless it is a technical issue. Performance metrics are recorded for quantitative data. After the completion of each task or all tasks, sometimes, interviews, semi-structured interviews, or group discussions are conducted for quality feedback. In the end, debriefing to participants of appreciation for their contributions is optional and can be delivered in an email.

Step 7: Reporting results

In academia, the results are usually reported in a poster, a presentation, a short paper, or a paper with an introduction of research background and research problems (objectives of testing), a literature review of previous studies on these problems, a presenting of the application and its design process, a designed experiment with materials, participants and procedure, results in analysis, discussion, limitations, and future directions.

To summarize data collected from the user's performance, a set of descriptive statistics can be applied : the counts of errors, the count of the correctness of each user's actions, the count of help, and the medium, mean or range of completion time, etc. To summarize data collected from questionnaires and interviews, for limited choice questions, we can compute the average score of each question for a large sample. For a small sample, this may not even be necessary in order to view trends. Also, obtaining the mean has a risk of losing some details. For open-ended questions and comments, list all questions and group all similar answers into meaningful categories.

Depending on the designed experiment, data can be compiled with groups, tasks, and versions of an application.

In order to develop possible improvements for the next iteration design process, more attention should pay to asset errors by identifying the occurrence probability of errors, sources of errors, and severity of errors.

5. Well-known Attributes and Measures across standards and models

Various standards and models list a range of attributes for usability. Each attribute reveals its real meaning only when the measure of the attribute is specified (Shackel, 1991). A summary of attributes and measures that are

highly cited is listed below (Table 1). These attributes and measures are developed from three main seminar work: Nielsen's five attributes (1994), ISO 9241-11(1993c) standards, and MUSiC (Bevan & Macleod, 1994).

Table 1: Attributes and Measures in standards and models

Usability Attributes	Definitions	Measures
Effectiveness (Bevan & Macleod, 1994; Seffah, Kececi, & Donyaee, 2006; Quesenbery, 2003; Shackel, 1991; Harrison, 2013; ISO 9241-11, 2018)	The required range of tasks must be accomplished at better than some required level of performance (e.g., in terms of speed and errors) by the target users in the usage environments (Shackel, 1991).	Each action time, kind and rate of errors, recovery from errors (Shackel, 1991).
	the extent to which the intended goals of use of the overall system are achieved (Bevan & Macleod, 1994)	The goal or sub-goals of using the system to the accuracy and completeness with which these goals can be achieved (Bevan & Macleod, 1994)
	The completeness and accuracy with which users achieve their goals (Quesenbery, 2003).	The accuracy of task completion and undetected errors (Quesenbery, 2003)
	The degree of accuracy and completeness with which the user achieves a specified task in a certain context. (Seffah et al., 2001)	Not mentioned directly
	The ability of a user to complete a task in a specified context (Harrison, 2013)	Task completion rate (Harrison, 2013)
	The accuracy and completeness with which users achieve specified goals. Lack of it can result in outcomes that could cause harm from use (ISO 9241-11, 2018)	-- Correctness of one task completed by one user -- Success rate of tasks completed by a group of users -- Error's frequency and other details -- User's perception of the correctness of one task completed by one user -- User's perception of the success rate of tasks completed by a group of users (ISO 9241-11, 2018)
Errors (Nielsen, 1994; Constantine and Lockwood, 1999; Harrison, 2013)	Any action that does not accomplish the desired goal (Nielsen, 1994)	Counting the number of such actions made by users while performing some tasks (Nielsen, 1994)
	(Reliability) leads its users to make fewer mistakes (Constantine and Lockwood, 1999)	the percent of completed work that was correct (Constantine and Lockwood, 1999)
	(Error Tolerant) How well the product prevents errors and helps the user recover from any that do occur (Quesenbery, 2003)	Observation of how easily or accurately users recover from problems (Quesenbery, 2003)
	how well the user can complete the desired tasks without errors (Harrison, 2013)	Count and evaluation of errors (Harrison, 2013)
Efficiency (Nielsen, 1994; Constantine and Lockwood, 1999; Bevan & Macleod, 1994; Seffah et al., 2006; Harrison, 2013; ISO 9241-11, 2018)	An expert user's steady-state level of performance after the initial learning phase completed (Nielsen, 1994)	Measure the time it takes an expert user to perform some typical test tasks (Nielsen, 1994)
	(Efficiency in use) Be efficient to use, leading to greater productivity on the part of its users (Constantine and Lockwood, 1999)	The percent correctly completed per unit time(Constantine and Lockwood, 1999)
	The resources such as time, money or mental effort that have to be spent to achieve the intended goals (Bevan & Macleod, 1994)	The time or resources a user spends on performing a task (Bevan & Macleod, 1994)
	The speed (with accuracy) with which this work can be done (Quesenbery, 2003)	Timing data, and subjective impressions from users (Quesenbery, 2003)
	The amount of resources expended in relation to the accuracy and completeness with which the user achieves a goal (Seffah et al., 2001)	Not mentioned directly

	The ability of the user to complete their task with speed and accuracy (Harrison, 2013)	The time to complete a given task, or the number of keystrokes required to complete a given task (Harrison, 2013)
	The resources(time, human effort, money, and materials) used in relation to the results achieved (ISO 9241-11 2018)	-- Time to complete the task -- Cost to complete the task -- User's perceived time to complete the task -- User's perceived cost to complete the task (ISO 9241-11 2018)
Learnability (Nielsen, 1994; Shackel, 2009; Constantine and Lockwood, 1999; Seffah et al., 2006; Harrison, 2013)	Initial ease of learning (Nielsen, 1994)	Measure the time it takes a novice user to reach a specified level of proficiency (complete a certain task successfully) in using it (Nielsen, 1994)
	The time to learn from commissioning based upon some specified amount of training and user support, and the time to relearn each time for intermittent users (same with casual users) (Shackel, 1991)	The time to learn (Shackel, 1991)
	(Rememberability) Easy to learn how to use (Constantine and Lockwood, 1999)	the percent of total assigned work completed within the allotted time (Constantine and Lockwood, 1999)
	How well the product supports both initial orientation and deeper learning (Quesenbery, 2003)	Time to learn (Quesenbery, 2003)
	The ease with which the features required for achieving particular goals can be mastered (Seffah et al., 2001)	Not mentioned from usability testing perspective
	The ease with which a user can gain proficiency with an application (Harrison, 2013)	How long it takes these participants to reach a pre-specified level of proficiency
Memorability (Nielsen, 1994; Constantine and Lockwood, 1999; Harrison, 2013)	Easy to remember for casual users (Nielsen, 1994)	Measure the time it takes a casual user to perform some typical test tasks after a period of time away from the system (Nielsen, 1994)
	Once learned, easy to remember how to use (Constantine and Lockwood, 1999)	Not mentioned directly
	The ability of a user to retain how to use an application effectively (Harrison, 2013)	Comparing the performance of similar tasks before and after a period of inactivity (Harrison, 2013)
Satisfaction (Nielsen, 1994; Constantine and Lockwood, 1999; Shackel, 2009; Bevan & Macleod, 1994; Seffah et al., 2006; Harrison, 2013;ISO 9241-11 2018)	Users should have an entertaining/moving/enriching experience when using such system (Nielsen, 1994)	Psychophysiological measures (if applicable), the user's subjective opinion from a questionnaire (Nielsen, 1994)
	Leaving them subjectively pleased about their experience using it (Constantine and Lockwood, 1999)	Not mentioned directly
	(Attitude) Acceptable levels of the human cost in terms of tiredness, discomfort, frustration and personal effort, so that satisfaction causes continued and enhanced the usage of the system (Shackel, 1991)	Not mentioned directly
	The extent to which the user finds the overall system acceptable (Bevan & Macleod, 1994)	Attitude rating scales such as SUMI, and indirectly measures, e.g. ratio of positive to negative comments during use, rate of absenteeism (Bevan & Macleod, 1994)
	(Engaging) How pleasant, satisfying or interesting an interface is to use (Quesenbery, 2003)	Interviews or surveys (Quesenbery, 2003)
	Freedom from discomfort and positive attitude towards the use of the software product (Seffah et al., 2001)	Not mentioned directly
	The perceived level of comfort and pleasantness afforded to the user through the use of the software and reflected in the attitudes of the user towards the software (Harrison, 2013)	Questionnaires and other qualitative techniques (Harrison, 2013)

	The extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet user's needs and expectations (ISO 9241-11, 2018)	-- Observed frequency of reuse -- Observed of certain emotions -- Satisfaction with experience (e.g. task achievement, trust, propensity to recommend, pleasure, comfort, etc.) (ISO 9241-11 2018)
Flexibility (Shackel, 1991)	Task flexibility (how a new product is integrated into people's work patterns) and environment flexibility (how a product performs in different environments)(Shackel, 1991) Not mentioned directly	
Adaptiveness (Bevan & Macleod, 1994)	It concerns the functionality of the device within a given application domain. (Measures are not mentioned.)	
Cognitive Load (Bevan & Macleod, 1994; Harrison, 2013)	the mental effort required to perform tasks and particularly important in safety-critical applications (Bevan & Macleod, 1994)	Heart rate variability, the Subjective Mental Effort Questionnaire (SEMQ) and the NASA Task Load Index (TLX) (Bevan & Macleod, 1994)
	The amount of cognitive processing required by the user to use the application. (Harrison, 2013)	NASA Task Load Index (TLX) (Harrison, 2013)
Productivity, Safety, Trustfulness, Accessibility, Universality, Usefulness (Seffah et al., 2006)	**Productivity** is the level of effectiveness achieved in relation to the resources (i.e. time to complete tasks, user efforts, materials or financial cost of usage) consumed by the users and the system. In contrast with efficiency, productivity concerns the amount of useful output that is obtained from user interaction with the software product. Macleod et al. (1997) noted that there are generally two types of user task one that is productive and the other is unproductive. The productive user task actions are those that contribute to the task output. Therefore, our definition of productivity considers the productive resources that are expended in accomplishing users' tasks. **Safety** concerns whether a software product limits the risk of harm to people or other resources, such as hardware or stored information. **Truthfulness** is the of faithfulness a software product offers to its users. It is most pertinent concerning e-commerce websites, it could potentially apply to many different kinds of software products. **Accessibility** is the capability of a software product to be used by persons with some type of disability. **Universality** concerns whether a software product accommodates a diversity of users with different cultural backgrounds. **Usefulness** is whether a software product enables users to solve real problems in an acceptable way. (Measures are not mentioned.)	

To summarize, the popular eight attributes used in summative usability studies include effectiveness, errors, efficiency, learnability, memorability, satisfaction, flexibility, cognitive workload. Measures for these attributes are categorized into five groups (Table 2).

Table 2: Attributes and grouped measures

Attributes	Grouped measures	
Effectiveness	Rate of task completion correctly and counts and details of errors	
Errors		
Efficiency	time and other resources spent on successful task completion	by all levels of users
Learnability		by novice users
Memorability		by medium-level users after a period of stopping
Flexibility	the count of incidences and time spent on unexpected actions in completing a task	
Satisfaction	psychophysiological measures (e.g. EEGs, pupil dilation, heart rate, blood pressure, etc); questionnaires (see Section 5.2), interviews, observed frequency of reuse and certain emotions	

6. Usability testing in Smartphone Application in retrieved papers

Mobile phone applications' usability principle has differences from other products on other platforms resulting from the distinct characteristics inherent in mobile phones and the particularities of the way people use it.

One distinct character is small screen size and only limited information can be displayed on one single screen. This causes some functions or useful information to be deeply hidden inside applications. Users with notorious short term memory will have difficulties to find the hidden information, let alone complete a task.

Another particularity is that unlike traditional desktop applications, users of mobile applications may be performing additional tasks, such as walking while using the mobile device. The multi-task mode of using a mobile phone increases the difficulties of mobile phone application's design. The PACMAD usability model (Harrison, 2013) includes cognitive workload as one attribute because users of mobile applications may be performing additional tasks, such as walking, watching tv, etc.while using the mobile device.

Usability testing for a mobile phone application needs to take these differences into consideration. More will be discussed in Section 8.

6.1. Attributes and measures in smartphone application usability testing

Attributes

Besides attributes previously mentioned in Table 1, some other attributes (Table 3) are used: usefulness and ease of use (Dantas et al., 2017; Reis et al., 2019; Hughes et al., 2019; Nazar & Zulfadli, 2017), feasibility (Casida et al, 2018), information quality (Hughes et al., 2019; Nazar & Zulfadli, 2017), user experience (Pratama et al., 2017; Hughes et al., 2019), functionality, reliability, maintainability (Yabut et al., 2017). Some papers adopt usability as a whole and do not analyze its attributes (Beatty et al., 2018; Ng, et al., 2017; Hughes et al., 2019; Wohlfahrt-Laymann et al., 2018; Saputra et al., etc.). These new attributes are explained briefly next.

The usefulness and ease of use (Dantas et al, 2017; Reis et al., 2019) are adopted from a questionnaire (USE). Hughes et al., (2019) does not provide the definitions of the usefulness and ease of use and evaluated by interviewing participants. The ease of use used in (Nazar & Zulfadli, 2017) is not clarified by the researchers, either. The researchers divide feasibility into three parts: acceptability, usability, and competency and designed a survey to evaluate them (Casida et al, 2018). Information quality is not given a further explanation(Hughes et al., 2019; Nazar & Zulfadli, 2017). Pratama et al., (2017) adopts the User Experience Questionnaire (UEQ) to evaluate user experience, and Hughes et al., (2019) does not explain the user experience and interviews the participants to capture their experience. Yabut et al,. (2017) select five attributes from ISO/IEC 9126 for software quality:

> **Functionality** - "A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs."
> **Usability** - "A set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users."
> **Efficiency** - "A set of attributes that bear on the relationship between the level of performance of the software and the number of resources used, under stated conditions."

Reliability - "A set of attributes that bear on the capability of the software to maintain its level of performance under stated conditions for a stated period of time."
Maintainability - "A set of attributes that bear on the effort needed to make specified modifications."

Table 3: Attributes and measures in smartphone application usability testing

Attributes	Measurements
Effectiveness	Correctness and errors (Birnstiel et al., 2019);
	Task (path movement) smoothness (Rodriguez, Dehghan, Figueroa, & Jagersand, 2018);
	Navigation time (completion time of each task), usage of features (captured by analyzing usage logs) (Kumar, Srivastava, Yadav, & Deshmukh, 2017);
	Task completed rate (Hashim & Lee, 2018; Baskoro & Widyanti, 2018; Adhy, Prasetio, Noranita, & Saputra, 2018; Zaror et al., 2019; Adli & Lestari, 2017;Tomaschko & Hohenwarter, 2018);
	Task successful per unit time (Nugraha, Syaifullah, & Puspasari, 2018);
	One section in a questionnaire with a 5-Likert Point (Hashim & Lee, 2018);
	Participant feedback (Kumar, Srivastava, Yadav, & Deshmukh, 2017);
	Observation of the accuracy level and time took to be efficiently accurate (Khan et al., 2017);
Errors	Most common users' errors, Obstacles of an interface to interaction(Garcia & de Lara, 2018);
	One part of a self-designed questionnaire with a 5-point Likert scale and two open-ended questions (Nurhudatiana, Hiu, & Ce, 2018);
Efficiency	Mode of the ratio between the number of interactions performed by the minimum quantity required to complete the task (Garcia & de Lara, 2018);
	Time-on-task and error percentage (Baskoro & Widyanti, 2018) ;
	Task completion time (Hashim & Lee, 2018; Rodriguez, Dehghan, Figueroa, & Jagersand, 2018; Zaror et al., 2019; Tomaschko & Hohenwarter, 2018);
	Task completed speed (Adhy, Prasetio, Noranita, & Saputra, 2018);
	Time to complete the task and the ease to find the intended page (Nugraha, Syaifullah, & Puspasari, 2018);
	Performance metrics(time on task, No. of ane deviation, No. of Collision, No. of Reverse) (Kumar, Srivastava, Yadav, & Deshmukh, 2017);
	One part of s self-designed questionnaire with a 5-point Likert scale and two open-ended questions (Nurhudatiana, Hiu, & Ce, 2018);
	ISO9126 evaluation tool with a 5-point Likert scale (Yabut et al., 2017)
Learnability	Average task execution time (Garcia & de Lara, 2018) ;
	Actual score/Ideal score of learnability (Adhy, Prasetio, Noranita, & Saputra, 2018);
	Average understanding failures (Garcia & de Lara, 2018) ;
	One part of a self-designed questionnaire with a 5-point Likert scale and two open-ended questions (Nurhudatiana, Hiu, & Ce, 2018);
	One part of a self-designed questionnaire with a 5-point Likert scale (Adli & Lestari, 2017);
	2 item in SUS (Ng et al., 2017)
Memorability	Average understanding failures (Garcia & de Lara, 2018);
	One part of a self-designed questionnaire with a 5-point Likert scale (Adli & Lestari, 2017);

Satisfaction/Enjoyment/Attractiveness	Usefulness, Satisfaction, and Ease of Use Questionnaire (USE) (Reis et al., 2019); Questionnaire for User Interface Satisfaction (QUIS) (Hernandez et al., 2019; Baskoro & Widyanti, 2018); Modified Mobile Application Rating Scale (M-MARS) (Quinn, Staub, Barr, & Gruber-Baldini, 2019; Woods, Duff, Roehrer, Walker, & Cummings, 2019) Single Ease Question (SEQ) 7-point Likert scale (Thyvalikakath, Schleyer and Monaco, 2007) System Usability Scale (SUS) (Zaror et al., 2019; Tomaschko & Hohenwarter, 2018); Quality in Use Integrated Measurement (QUIM) (Wardhana, Sabariah, Effendy, & Kusumo, 2017); A self-designed questionnaire (Hashim & Lee, 2018; Wichiennit, Sunat, Chiewchanwattana, Louchaisa, & Onnoom, 2017; Lee & Kim, 2019; Adli & Lestari, 2017; Shada & Ayu, 2018; Nurhudatiana, Hiu, & Ce, 2018); Rating scale for user's satisfaction with function and characteristics (Nugraha, Syaifullah, & Puspasari, 2018); questionnaires, structure and semi-structured interviews and participant observations(Sumanasekera, Mihilar, Wickramasinghe, & Arunathilake, 2018); Actual score/Ideal score of satisfaction (Adhy, Prasetio, Noranita, & Saputra, 2018);
Task Cognitive Load	NASA-TLX survey (Rodriguez, Dehghan, Figueroa, & Jagersand, 2018);
Usability	Task completion rate (Beatty, Magnusson, Fortney, Sayre, & Whooley, 2018; Ng, Cheong, Hajimohammadhosseinmemar, & Yap, 2017); System Usability Scale (SUS) (Hughes et al., 2019; Wohlfahrt-Laymann, Hermens, Villalonga, Vollenbroek-Hutten, & Banos, 2018;Saputra, Farhan, & Irvanizam, 2018; Wichiennit, Sunat, Chiewchanwattana, Louchaisa, & Onnoom, 2017; Pratama, Setiawan, & Wibirama, 2017; Zaror et al., 2019; Veale, Dogan, & Murphy, 2019; Paldán et al., 2019; Beatty, Magnusson, Fortney, Sayre, & Whooley, 2018; Ahmed, Abubakar, Ibrahim, Garry, & Andrew, 2018;Quinn, Staub, Barr, & Gruber-Baldini, 2019; Ng, Cheong, Hajimohammadhosseinmemar, & Yap, 2017; Post-Study System Usability Questionnaire (PSSUQ) (Crosby et al., 2017); ISO9126 evaluation tool with a 5-point Likert scale （Yabut et al., 2017) A self-designed questionnaire (Ding et al., 2017, Mallan, Gopi, Muir, & Bhavani, 2017; Alkhafaji, Cocea, Crellin, & Fallahkhair, 2017; Choo et al., 2017; Beatty, Magnusson, Fortney, Sayre, & Whooley, 2018; Aragon, Castillo, Agustin, & Aguilar, 2019; Shada & Ayu, 2018); Comments from think-aloud principle (Choo et al., 2017; Veale, Dogan & Murphy, 2019) Comments from cognitive walkthrough and semi-structured interview (Veale, Dogan & Murphy, 2019)
Usefulness and ease of use	Usefulness, Satisfaction, and Ease of Use Questionnaire (USE) (Dantas et al, 2017; Reis et al., 2019); Semi-structured questions(Hughes et al., 2019);
Ease of use	A survey adopted from other papers (Nazar & Zulfadli, 2017);
Feasibility (acceptability, usability, and competency)	A self-designed questionnaire with a 5-point Likert scale, Semi-structured, face-to-face interviews for acceptability, usability, and competency (Casida, Aikens, Craddock, Aldrich, & Pagani, 2018)
Information quality	Semi-structured questions (Hughes et al., 2019); A survey based on other papers (Nazar & Zulfadli, 2017)
User experience	User Experience Questionnaire (UEQ) (Pratama, Setiawan, & Wibirama, 2017); Semi-structured questions(Hughes et al., 2019);
Functionality, Reliability, Maintainability	ISO9126 evaluation tool with a 5-point Likert scale (Yabut et al., 2017)

Measures

Various measures were adopted in the retrieved papers. Effectiveness/errors and efficiency are consistently measured: for effectiveness, the success rate of tasks and error counts are routinely used, and for efficiency, various time-related measures such as time spent on a task are used. For satisfaction, a wider range of measures provided via a number of standard and pre-validated questionnaires are applied in papers: Usefulness, Satisfaction, and Ease of Use Questionnaire (USE) (Reis et al., 2019); Questionnaire for User Interface Satisfaction (QUIS) (Hernandez et al., 2019; Baskoro & Widyanti, 2018); Modified MobileApplication Rating Scale (M-MARS) (Quinn et al., 2019; Woods et al., 2019); Single Ease Question (SEQ) (Thyvalikakath et al., 2007); System Usability Scale (SUS) (Zaror et al., 2019; Tomaschko & Hohenwarter, 2018, etc.); Quality in Use Integrated Measurement (QUIM) (Wardhana, Sabariah, Effendy, & Kusumo, 2017). Meanwhile, other studies (Hashim & Lee, 2018; Wichiennit et al., 2017; Lee & Kim, 2019; Adli & Lestari, 2017; Shada & Ayu, 2018; Nurhudatiana et al., 2018) develop their own questionnaires for satisfaction. Interviews, semi-structured interviews, focus groups, observations are also used by researchers (Hashim & Lee, 2018; Sumanasekera et al., 2018; Khan et al., 2017). A study (Adhy et al., 2018) formed an equation to analyse survey scales for satisfaction.

*Ideal score =biggest score scale * number of respondents*
Actual score = score of each question in the questionnaires
Percent = Actual ScoreIdeal Score x 100%

It is evident that a wide range of measures have been used to assess "satisfaction", and these measures also overlap with the measures used for the effectiveness and efficiency aspect of the software. More details and discussions will be provided in Section 7 on Satisfaction.

Questionnaires

In this section, we will describe the questionnaires chronologically mentioned in Tables 1 and 3.

Questionnaire for User Interface Satisfaction (QUIS) (Chin et al., 1988) (Citation count: 1865) (See Appendix A)

The Questionnaire for User Interaction Satisfaction (QUIS) was developed to evaluate the user's subjective responses with a computer's interface. It has also been applied on other products. It consists of a demographic questionnaire, and overall reaction to the satisfaction, and subjective responses to the specific system components: "screen"; "terminology and system information"; "learning"; "system capability"; "usability and UI". (See https://www.cs.umd.edu/hcil/quis/)

NASA-Task Load Index (TLX) survey (Hart & Staveland, 1988) (Citation count: 9155) (Appendix B)

The NASA Task Load Index (NASA-TLX) is a widely-used, subjective, multidimensional assessment tool that rates perceived workload in order to assess a task, system, or team's effectiveness or other aspects of performance. The total workload is divided into six subjective attributes:

> **Mental Demand** -- How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
>
> **Physical Demand** -- How much physical activity was required? Was the task easy or demanding, slack or strenuous?
>
> **Temporal Demand** -- How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?

Overall Performance -- How successful were you in performing the task? How satisfied were you with your performance?
Effort -- How hard did you have to work (mentally and physically) to accomplish your level of performance?)
Frustration -- How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

These are rated for each task on a 20-point scale.

Software Usability Measurement Inventory (SUMI) (Citation count: 655) (Kirakowski & Corbett, 1993) (Appendix C)
The concept of usability as assessed by SUMI draws on the definition in ISO 9241-11 (1993). It consists of five attributes: effect, efficiency, learnability, helpfulness, and control to measure usability. All statements have a 3-point Likert-scale: agree, undecided, disagree. The definition of each attribute is available in the SUMI manual, yet not accessible. In book *Usability Evaluation In Industry* (Jordan, Thomas, McClelland, & Weerdmeester, 1996), their definitions are listed. **Affect** is the user's general emotional reaction to the software. **Efficiency** measures the degree to which users feel that the software assists them in their work and it is related to the concept of transparency. **Learnability** measures the speed and facility with which the user feels that he or she has been able to master the system, or to learn how to use new features when necessary. **Helpfulness** dimension measures the degree to which the software is self-explanatory with the adequacy of help facilities and documentation. Finally, **control** measures the extent to which the user feels in control of the software when carrying out the task. (See: http://sumi.uxp.ie/)

The Usefulness, Satisfaction, and Ease of use (USE) (Lund, 2001) (Citation count: 856) (Appendix D)
It measures usability from usefulness, ease of use, ease of learning, and satisfaction with a seven-Likert scale, ranging from strongly disagree to strongly agree.

Post-Study System Usability Questionnaire (PSSUQ) (Lewis,1995; 2002)(Citation count: 2189; 386) (Appendix E)
The Post-Study System Usability Questionnaire (PSSUQ) is currently a 19-item instrument for "assessing user **satisfaction** with system usability". The items are 7-point graphic scales, anchored at the endpoints with strongly agree (1) and strongly disagree (7) and a not applicable (N/A) point outside the scale. The items assess four aspects of usability:
System usefulness – calculated by taking the average of questions 1-8,
Information quality – calculated by taking the average of questions 9-15,
Interface quality – calculated by taking the average of questions 16-18,
Overall user satisfaction – calculated by taking the average of questions 19.

System Usability Scale (SUS) (Brooke, 1996) (Citation count: 8101) (Appendix F)
Brooke (1996) emphasized the context and used the definition of usability, IS0 9241-11(1995). The System Usability Scale (SUS) covers a variety of aspects of system usability, such as the need for support, training, and complexity. (See https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html)

User Experience Questionnaire (UEQ) (Laugwitz, Held, & Schrepp, 2008) (Citation count: 704) (Appendix G)

The questionnaire covers a comprehensive impression of user experience. Both classical usability attributes (efficiency, perspicuity, dependability) and user experience attributes (activeness, stimulation, novelty) are measured. All statements have a 7-Likert scale.
Efficiency -- Can users solve their tasks without unnecessary effort? Does it react fast?
Perspicuity -- Is it easy to get familiar with the product and to learn how to use it?
Dependability -- Does the user feel in control of the interaction? Is it secure and predictable?
Activeness -- Overall impression of the product. Do users like or dislike it?
Stimulation -- Is it exciting and motivating to use the product? Is it fun to use?
Novelty -- Is the design of the product creative? Does it catch the interest of users?
(See: https://www.ueq-online.org/)

Mobile App Rating Scale (MARS) (Stoyanov et al., 2015) (Cited by 537)(Appendix H)
The Mobile App Rating Scale (MARS) is a well-known standardized tool developed by the Queensland University of Technology for health apps' comparison. It is designed to score apps on the attributes of engagement, functionality, aesthetics, and information quality:
Engagement -- including sub-attributes: fun, interesting, customizable, interactive (eg, sends alerts, messages, reminders, feedback, enables sharing), and well-targeted to the audience, each sub-attribute is assessed with a set of questions/statements.
Functionality -- including sub-attributes: app functioning, easy to learn, navigation, flow logic, and gestural design of the app, each sub-attribute is assessed with a set of questions/statements.
Aesthetics -- graphic design, overall visual appeal, color scheme, consistent style,
Information -- Contains high-quality information (eg, text, feedback, measures, references) from a credible source.
And all statements have multi-level scales.

6.2. Usability testing design in smartphone application usability testing

Settings

Compared with products on other platforms, e.g. desktops, laptops, and VR glasses, etc., one unique characteristic of mobile phone application usability testing is that it can be conducted in the field setting due to the accessibility of mobile phone, while it can be used in a laboratory setting with controls. Nine papers conducted usability testing in a laboratory, whereas thirteen papers operated their usability testing in the field. The rest of papers do not mention explicitly. The participants in a study of mobile support for older adults and their caregivers (Quinn, et al., 2019) used the application in their daily work and life for a one-month period. Meanwhile, another health care application usability testing, an upper extremity stroke rehabilitation application, was conducted in the lab setting (Hughes et al., 2019).

Table 4: Settings in application usability testing

	Papers
Laboratory setting	Birnstiel et al., 2019; Hughes et al., 2019; Baskoro & Widyanti, 2018; Rodriguez, Dehghan, Figueroa, & Jagersand, 2018; Adhy, Prasetio, Noranita, & Saputra, 2018; Mallan, Gopi, Muir, & Bhavani, 2017; Tomaschko & Hohenwarter, 2018; Beatty, Magnusson, Fortney, Sayre, & Whooley, 2018; Choo et al., 2017
Field setting	Hernandez et al., 2019; Baskoro & Widyanti, 2018; Sumanasekera, Mihilar, Wickramasinghe, & Arunathilake, 2018; Alkhafaji, Cocea, Crellin, & Fallahkhair, 2017; Woznowski, Burrows, Laskowski, Tonkin, & Craddock, 2017; Khan et al., 2017; Ng et al., 2017; Quinn, Staub, Barr, & Gruber-Baldini, 2019; Woods, Duff, Roehrer, Walker, & Cummings, 2019; Zaror et al., 2019; Veale, Dogan, & Murphy, 2019; Paldán et al., 2019; Casida, Aikens, Craddock, Aldrich, & Pagani, 2018
Not mentioned	Reis et al., 2019; Garcia & de Lara, 2018; Hashim & Lee, 2018; Nurhudatiana, Hiu, & Ce, 2018; Wohlfahrt-Laymann, Hermens, Villalonga, Vollenbroek-Hutten, & Banos, 2018; Shada & Ayu, 2018; Saputra, Farhan, & Irvanizam, 2018; Nugraha, Syaifullah, &

Puspasari, 2018; Ding et al., 2017; Kumar, Srivastava, Yadav, & Deshmukh, 2017; Yabut et al., 2017; Nazar & Zulfadli, 2017; Wardhana, Sabariah, Effendy, & Kusumo, 2017; Wichiennit, Sunat, Chiewchanwattana, Louchaisa, & Onnoom, 2017; Adli & Lestari, 2017; Pratama, Setiawan, & Wibirama, 2017; Aragon, Castillo, Agustin, & Aguilar, 2019; Ahmed, Abubakar, Ibrahim, Garry, & Andrew, 2018; Crosby et al., 2017

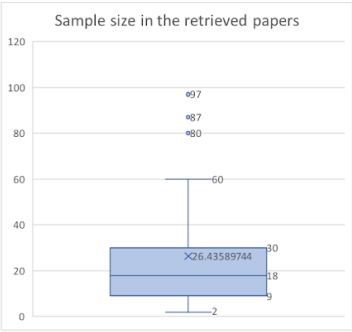
The laboratory setting and field setting have their own pros and cons respectively. Zhang and Adipat (2005) points out that the laboratory setting has three advantages. First, researchers have full control of usability testing. Secondly, fewer uncontrollable variables lead to easy-to-measure attributes and clean data. Thirdly, video or audio recordings are possible in the laboratory. On the other hand, they suggest the field setting leads to a maximum approach to the practical context users using a mobile phone application. Furthermore, they claim measuring usability in the field is far from trivial based on some challenges the field setting faces. Sufficient control of participants and the environment is lacking in general in the fielding testing. While fields may provide a more realistic setting, they nonetheless are not capable of fully capturing the complexity presented in the diverse usage scenarios in real life. Moreover, some data collection methods cannot be applied in the field, such as observation, think-aloud for long experiment sessions.

Notwithstanding those issues the field setting has, its advantage is incomparable and unexampled. Visually impaired people tested a botanical garden tour application's usability by actually experiencing the beauty of the garden with the App in the garden (Birnstiel et al., 2019). A 30-day long usability testing was conducted for a self-management application in patients and caregivers' regimen daily. Identifying suitable settings for a usability study is fundamental, as written in usability's definition, *"the extent to which a system, product, service can be used by specified users … in* ***a specified context of use*** *."* (ISO 9241-11, 2018).

Now we are seemingly stuck in a dilemma between the two settings, some studies have proposed a middle ground for usability testing. A controlling application of a robot was tested in a simulation environment to reduce the influences of environment issues (Rodriguez, et al., 2018). Similarly, a usability testing of a healthcare application was performed in an environment analogous to a typical clinic office in a usability research center (Choo et al., 2017).

Meanwhile, VR-based Softwares have been used broadly to study eating disorders (Pla-Sanjuanelo et al., 2017, 2019, Ferrer-Garcia et al., 2017, 2019; Paslakis et al., 2017). Specifically, participants in Pla-Sanjuanelo's study (2017) wore VR goggles, watched a jogging video, and tried to engage in simple movements. From these studies, we can see the potency of VR or AR being applied to usability testing. In other words, technology and creativity can combine these two settings to get the best of both. In any case, identifying a suitable setting for a usability study is fundamental, as written in usability's definition, *"the extent to which a system, product, service can be used by specified users … in* ***a specified context of use*** *."* (ISO 9241-11, 2018).

User group/Sampling, Task scenarios, Applications, Experimental design



The majority of the experiments used convenience sampling, within-group design, multiple tasks (up to 7 tasks), one application (or one version of the application), ranging from 2 to 98 (median: 18) participants (Figure 4).

Figure 4: Sample size in the retrieved papers

Devices

Two studies (Baskoro & Widyanti, 2018; Tomaschko & Hohenwarter, 2018) used eye-tracking devices. The setup is shown in Image 1.

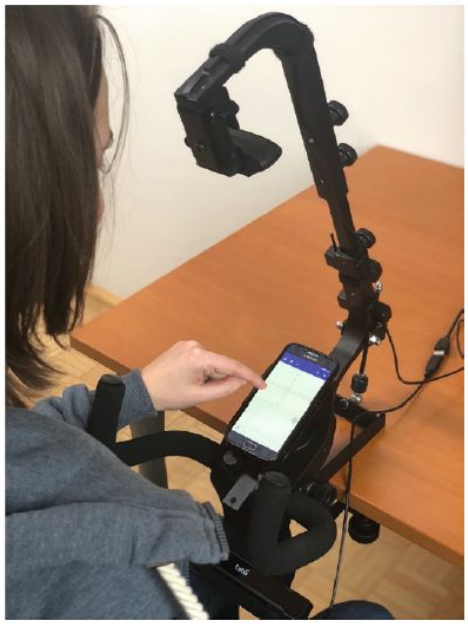


Image 1：A participant is using a smartphone with an eye-tracking device recording her/his operations (Tomaschko & Hohenwarter, 2018).

6.3. Research ethics and policy in smartphone application usability testing

Summative usability studies with the intention to generalize the findings are considered as human subject research.

IRB approval certifies that usability testing follows the basic ethical rules and presents no harm to participants. A signed informed-consent form indicates that participants understand how the experiment will be conducted and associated risk and/or benefit, and they agree to participate in the study. IRB procedure is a standard practice in the Western World, but not all countries or journals follow the best practice related to the procedure and requirements. Among retrieved articles, only 25% (Table 5) mentioned the obtain of IRB approval or/and signed a consent form. Two papers (Khan et al., 2017; Adhy, Prasetio, Noranita, & Saputra, 2018) stand out, regrettably, by presenting the participants' real names in the paper.

Table 5: Research ethics and policy in smartphone application usability testing in retrieved papers

	Papers
IRB approval or/and Signed consent form	Hughes et al., 2019; Hernandez et al., 2019; Garcia & de Lara, 2018;Baskoro & Widyanti, 2018; Sumanasekera, Mihilar, Wickramasinghe, & Arunathilake, 2018; Quinn, Staub, Barr, & Gruber-Baldini, 2019; Zaror et al., 2019; Paldán et al., 2019; Beatty, Magnusson, Fortney, Sayre, & Whooley, 2018; Choo et al., 2017; Crosby et al., 2017

6.4. Data analysis methods in smartphone application usability testing

Counts, mean with standard, T-test has been adopted by the majority of retrieved papers to analyze counts of task completion, counts of errors, time spent on task completion, and scores collected through questionnaires. For ordinal data collected through Likert scales, means are also used. Some papers also employ some testing methods for validation, for example, Pearson correlation and Cronbach's alpha is calculated to test QUIS's validation and reliability (Baskoro & Widyanti, 2018). The Shapiro-Wilk test is performed to confirm the SUS scores are normal distribution in Wohlfahrt-Laymann et al.(2018).

For qualitative data (e.g., interviews, answers to open-ended questions), all papers use content analytic procedures. In addition, some papers mention the methods of distilling information, for example, grounded theory analysis and affinity diagramming (Hernandez et al., 2019), and Braun & Clarke's process (Woods et al., 2019). Grounded theory (GT) is a systematic methodology used in the social sciences involving the construction of a theoretical model through four stages: codes, concepts, categories, and theory (Martin & Turner, 1986). Affinity diagramming means organizing related facts into distinct clusters. It is also known as affinity mapping, collaborative sorting, snowballing, or sometimes card sorting (Brassard, 1991; Scupin, 1997). Braun & Clarke's process () is a 6-step procedure, consisting of familiarization of the data through re-reading the transcripts (Step 1), generation of initial codes and writing them directly on the transcript segments considered interesting or meaningful to the analyst (Step 2), organization of codes into potential themes (Step 3), review of themes through checking and generating a thematic "map" (Step 4), generation of clear definitions and names for each theme (Step 5), and production of the report with compelling examples through a final analysis (Step 6) (Braun & Clarke, 2006).

7. The Satisfaction attribute and it measures

Satisfaction as one of the main factors in usability studies is less well defined as compared to effectiveness and efficiency attributes. Its measures are more diverse and nuanced. Further, the concept of satisfaction has also drawn attention from a number of other fields, including marketing, psychology, information systems, etc. A number of definitions for "satisfaction" proposed in HCI and other fields are presented and discussed below.

7.1. The concept of "satisfaction" in HCI

Yuksel and Yuksel (2001a) points out that no consensus is reached on definitions of satisfaction in the literature for customer satisfaction of hospitality and tourism. Similar in the HCI field, satisfaction has various definitions in standards and models of usability. In the following, we attempt to organize various thoughts on satisfaction along the dimensions proposed by ISO 9241-11(2018), and Jones and Suh (2000). (Table 6)

In ISO 9241-11(2018), satisfaction is the user's physical, cognitive and emotional responses that result from the use of a product to meet user's needs and expectations. The components of satisfaction that are important will depend on the reasons for considering usability. **Physical responses** are feelings of comfort or discomfort that represent physical components of satisfaction. They result from the physical experience of using the object of interest. **Cognitive responses** imply attitudes, preferences, and perceptions that represent cognitive components of satisfaction. Attitudes and perceptions can include trust, perceived degree of safety, perceived degree of security, and perceived extent of privacy. They result from the experience of use of the object of interest and can also be influenced by the experience of using similar systems and by other people's opinions. **Emotional responses** represent affective components of satisfaction. They result from experience while using the object of interest. These responses can be influenced by the experience of using similar systems and other people's opinions. Emotional responses can be assessed by physiological responses such as skin conductance, facial expression, as well as by self-assessment using rating scales.

Shackel (1991) uses "attitude" for user's satisfaction, and refers to the human cost in terms of tiredness, discomfort, frustration, and personal effort while using a product.

Among five attributes Nielsen (1994) proposed, satisfaction means how pleasant it is for the user to use the system. He points out satisfaction here is different from public general attitudes towards a platform, e.g. computers, mobile phones, or VR glasses, etc.even though it is likely that a person's feelings towards the platforms will impact user satisfaction. Additionally, he emphasizes that user's subjective ratings of its difficulty are much more closely related to peak difficulty they experienced than to mean difficulty.

Constantine and Lockwood (1999) define satisfaction as "software that satisfies users, leaving them subjectively pleased about their experience using it, is more useful than software that irritates or displeases users."

Bevan and Macleod (1994) suggest that satisfaction describes the perceived usability of the overall system by its users and the acceptability of the system to the people who use it and to other people affected by its use.

Engaging as one attribute in the 5E model refers to how pleasant, satisfying or interesting an interface is to a user (Quesenbery, 2003), and can be considered as "satisfaction" in this context.

The QUIM model (Seffah et al., 2006) defines satisfaction as the subjective responses from users about their feelings when using the software.

A debating on satisfaction definition reviewed in Yuksel and Yuksel (2001a) is about whether satisfaction is a cognitive evaluation or an emotional state, or the link between both the cognitive and emotional processes because customer satisfaction is an emotional feeling in response to a process of confirmation and/or disconfirmation (cognitive). In addition, ISO 9241-11 extends satisfaction's scope to physical responses of comfort and discomfort.

Some researchers suggest that customer satisfaction can be defined at least at two levels: transaction-specific satisfaction and overall satisfaction (Bitner & Hubbert, 1994). Transaction-specific dis/satisfaction only concerns a discrete service encounter, while overall dis/satisfaction refers to an overall assessment of all experience with service. More importantly, consumers view these two conceptualizations of satisfaction differently (Jones and Suh, 2000).

Transaction-specific satisfaction in Jones and Suh (2000) refers to user's impression resulted from one time encounter with the system (hospitality and tourism experiences), but it could be generalized to differentiate user's satisfaction with aspects of a product and user's overall satisfaction with the interaction with the entire product, see also QUIS, or user's satisfaction of one time experience with a product and user's satisfaction with multiple times experience with a product.

Combining the frameworks presented above, we can categorize the other conceptualization of satisfaction presented in other studies along the dimension of cognitive, emotional, and physical, the dimensions of aspect-specific and overall satisfaction, and the dimension of aspects-specific and entire product.

Table 6: Satisfaction's definition mapping on three spectrums

Satisfaction in Usability	Cognitive or Emotional or Physical	One time or Multiple times	Aspects-specific or Entire product
Shackel, 1991	All	Both	Both
Nielsen, 1994	Emotional	Both	Both
Constantine and Lockwood, 1999	Emotional	Both	Both

Bevan and Macleod, 1994	Cognitive	Both	Entire
Quesenbery, 2003	Emotional	Both	Both
Seffah et al., 2006	Emotional	Both	Both
ISO 9241-11, 2018	All	Both	Both

Table 6 shows a categorization of the definitions of satisfaction used in usability, based on those three spectrums of satisfaction. The scope of satisfaction as one attribute in ISO 9241-11 (2018) is the most comprehensive, covering emotional and cognitive states and assessing one-time interaction and overall experience of a product.

The discovery that the definitions of satisfaction in usability are varied makes more literature review of satisfaction imperative. Next, we will review the research on satisfaction in other areas.

7.2. Satisfaction in Psychology --- Self - Determination Theory

Self - Determination Theory (SDT) is initially proposed and then developed by Deci and Ryan (Deci & Ryan, 1985, 2000; Ryan & Deci, 2008). These authors recognized that needs to specify innate psychological nutriments essential for ongoing psychological growth, integrity, and well being, and they identified the three needs:the needs for competence, relatedness, and autonomy. **Competence** refers to the need to feel capable and effective in one's actions. **Relatedness** involves the need for belonging, intimacy, and connectedness to others. **Autonomy** indicates the need to experience one's behavior as freely chosen and volitional, rather than imposed by external forces. Not only psychological development and well-being but also goal-directed behavior cannot be achieved without addressing the needs that influence the process of pursuing goals. These three needs are essential for understanding the *what* (i.e., content) and *why* (i.e., process) of goal pursuits.

A product, as one part of an environment, only when it has the capacity to gratify one or all these three needs, will enhance the user's intrinsic motivation. If the intrinsic motivation of the user was not invoked and supported by the interface design, they would be less likely to get engaged in learning and using the application. For instance, if people did not experience satisfaction from learning how to learn a system for its own sake, but instead needed to be prompted by external reinforcements, they would be less likely to get engaged in learning new skills.

7.3. Satisfaction in Psychology -- Two-factor theory (motivator-hygiene theory)

Herzberg (1964) analyzed interviews with 203 engineers and accountants about their extremely happy or unhappy period with their jobs. He found that the nature of the work that people perform in their jobs has the capability to fulfill some needs such as achievement, competency, self-realization, personal worth, and social status, and the fulfillment results in happiness and satisfaction. However, the absence of gratified needs does not lead to unhappy or dissatisfied feeling/status. Conversely, dissatisfaction comes from unpleasant factors in jobs, e.g. annoying supervisor, terrible working conditions, complicated interpersonal relations.

Two-factor theory categories factors that impact job satisfaction into two slots: motivators and hygiene factors. Motivators deliver positive satisfaction, rooted in intrinsic conditions of the job itself. They can be challenging but meaningful work, a sense of responsibility to an organization, recognition of achievement, etc. Hygiene factors (e.g. job security, salary, bonus, working conditions, health insurance, paid vacations) do not give positive satisfaction or lead to higher motivation, though dissatisfaction results from their absence.

Motivators and hygiene factors in usability were not discussed in the publications reviewed in this article, and it is unclear if such factors have been identified. This could be an interesting direction to explore in future research.

7.4. Satisfaction in Information -- Information System Success Model

The information systems success model (IS success model) aims to offer a comprehensive understanding of IS success by positing six major dimensions or categories of IS success -- system quality, information quality, use, user satisfaction, individual impact, and organizational impact. It was initially developed by DeLone and McLean in 1992 and further update by them in 2003 (DeLone & McLean, 1992, 2003). User satisfaction is the recipient responds to the use of the output of an information system (DeLone & McLean, 1992). They recognize a key issue that whose satisfaction should be measured, sales representatives, executives, or managers. Studies have found that user satisfaction is associated with user attitudes toward computer systems (Igerhseim 1976; Lucas 1978), which is mentioned in Nielsen's book (1994). Goodhue (1986) suggests "information satisfactoriness" as an antecedent to and surrogate for user satisfaction. Information satisfactoriness is defined as the degree of match between task characteristics and information system functionality. This concept is quite close to utility, which is whether the system does what is needed functionally (Nielsen, 1993).

In the IS success model, quality has three major dimensions: "information quality," "systems quality," and "service quality." Each of them singularly or jointly affects "user satisfaction" (DeLone & McLean, 2003) (Figure 5).

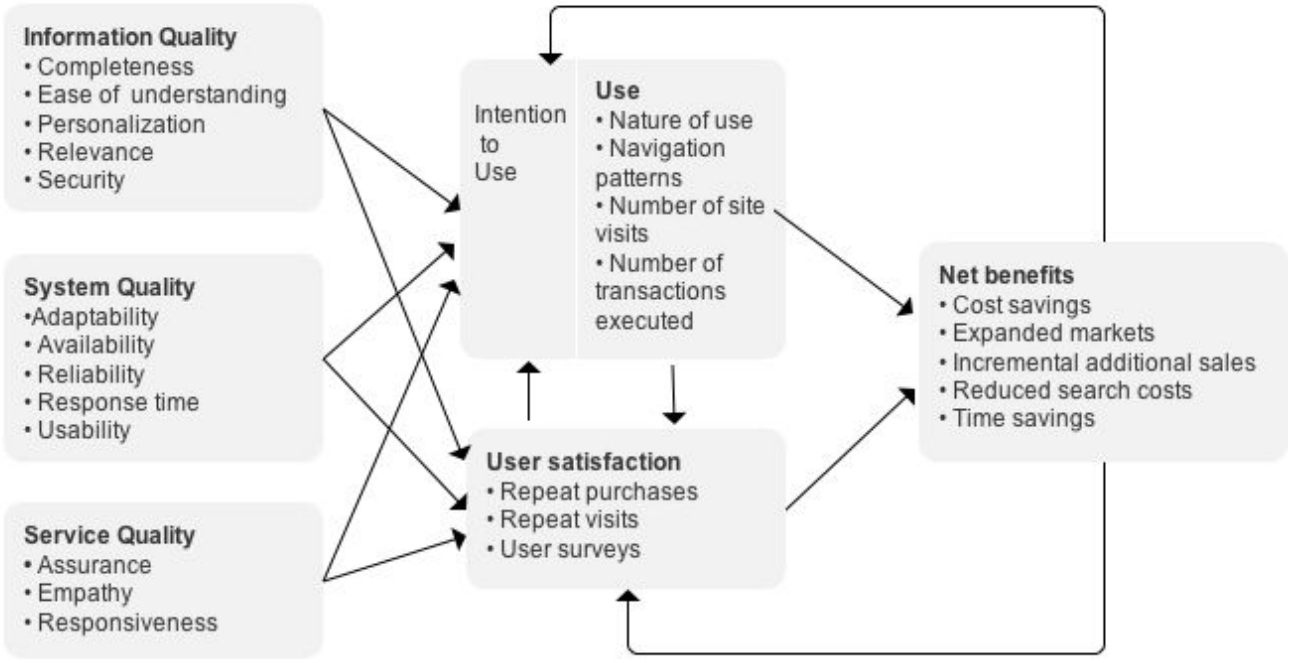


Figure 5: D&M IS Success Model, adopted from DeLone & McLean, 2003

The IS success model has been applied to measure e-commerce success (DeLone & McLean, 2003, 2004). "System quality" captures the characteristics of an e-commerce system: usability, availability, reliability,

adaptability, etc. "Information quality" measures e-commerce content quality. Web content should be complete, relevant, easy to understand, and secure. "Service quality" refers to the overall support delivered by the service provider. "Use" measures everything from a visit to a Web site to navigation within the site to information retrieval to the execution of a transaction. "User satisfaction" remains an important means of measuring customers' opinions of an e-commerce system and should cover the entire customer experience cycle from information retrieval through purchase, payment, receipt, and service. "Net benefits" are the most important success measures as they capture the balance of positive and negative impacts of e-commerce on our customers, suppliers, employees, organizations, markets, industries, economies, and even our societies.

7.5. Satisfaction in e-Commerce -- Expectancy-Disconfirmation Theory and Mobile catering application success model

A considerable attention has been paid to customer satisfaction in hospitality and tourism field (Oliver, 1980, 1992, 1993; Westbrook, 1980; Westbrook & Oliver, 1991; Peterson & Wilson, 1992; Walker, 1995; Dabholkar, Shepherd, & Thorpe, 2000; Yuksel & Yuksel, 2001), and one popular theory, the Expectancy-Disconfirmation Theory (Oliver, 1980, 1992, 1993), has been applied to e-commerce and mobile phone applications. It is a cognitive theory which seeks to explain post-purchase or post-adoption satisfaction as a function of expectations, perceived performance, and disconfirmation of beliefs.

Szymanski and Hise (2000) initially examined the factors that make customers satisfied with their e-retailing experiences, and discovered the convenience, site design, and financial security are strong predictive factors on consumer judgment of e-satisfaction. Soon after that, a measurement of web-customer satisfaction for one phrase, where customers searching for their target products, was developed (McKinney, Yoon, & Zahedi, 2002). They synthesizing the Expectation-Disconfirmation theory divided website quality into information quality (IQ) and system quality (SQ) and then presented nine factors for website customer satisfaction. The nine factors are relevance, timeliness, reliability, scope, perceived usefulness, access, usability, navigation, and interactivity. This measurement model was tested and proved to have a high degree of validity and reliability. The subsequent investigation (Wolfinbarger & Gilly, 2003) suggested four factors—website design, fulfillment/reliability, privacy/security, and customer service—are the domain factors in customer assessment of quality and satisfaction, customer loyalty and attitudes toward the retail website.

EDT has been adopted and used in a large number of customer satisfaction research, e.g., Bai, Law, and Wen (2008) and (Wang, Tseng, Wang, Shih, & Chan, 2019). Wang et al. (2019) proposed a mobile catering application success model on the foundation of the IS success model (DeLone & McLean, 1992, 2003), and the IS success model in the context of e-commerce (DeLone & McLean, 2004), product marketing theories, product quality research, cognitive evaluation theory (Ryan and Deci, 2000), and the Expectancy - Disconfirmation theory (Oliver, 1980). After collecting sample data using an online survey, partial least squares structural equation modeling (PLS-SEM) is selected for analysis to verify the model and clarify the inner relationships among each component (Figure 6).

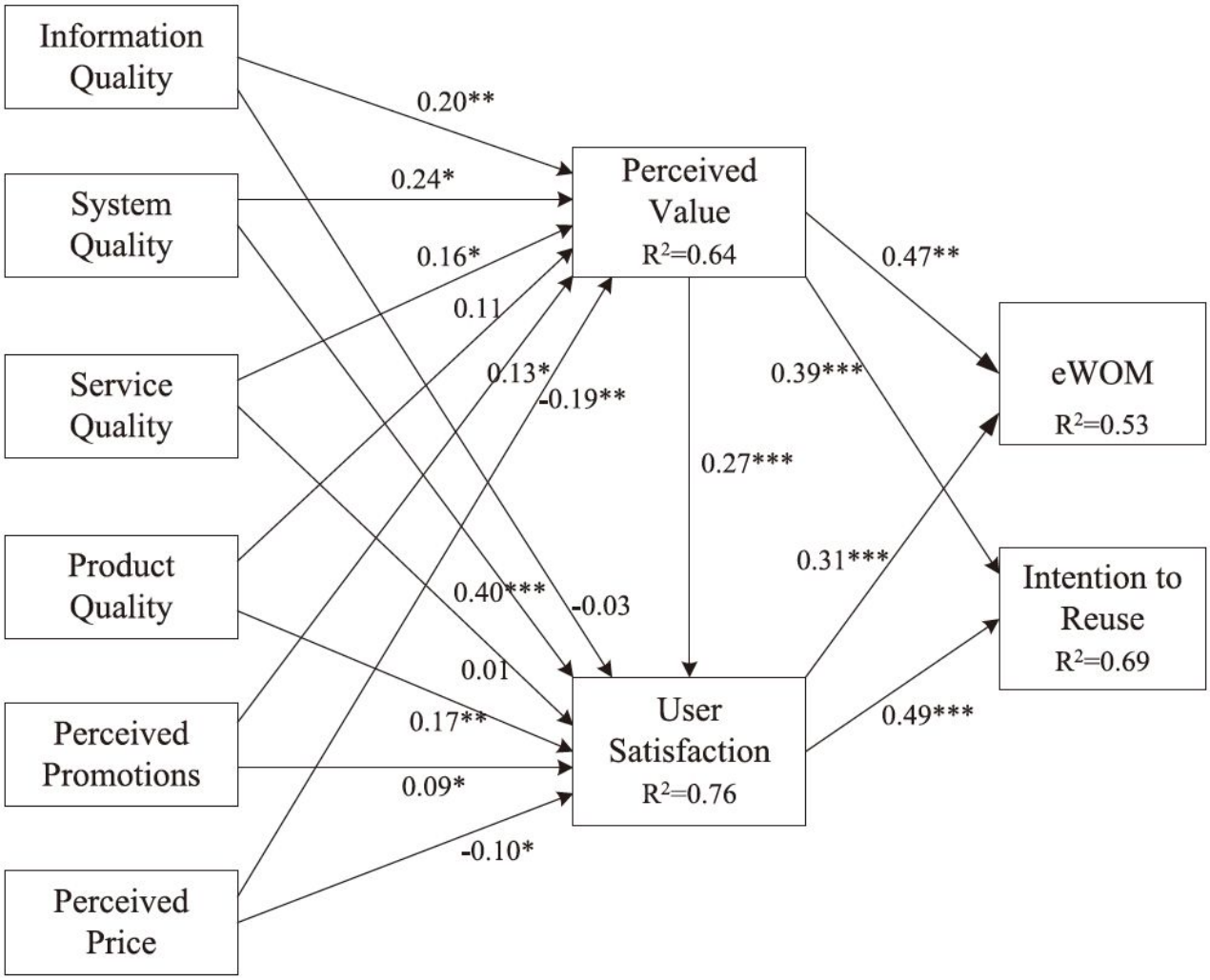


Figure 6: A Note: A mobile catering application success model. EWOM stands for electronic word-to-mouth. Standardized path coefficients are shown. *p < 0.05; **p < 0.01; ***p < 0.001.

Usability is covered in information quality and system quality (Figure 5). Satisfaction is affected by system quality, product quality, perceived promotions, and perceived value positively, and influenced by perceived price negatively.

8. General Discussion

The relationships among acceptability, usefulness, utility, usability, and user experience have been elaborated in Section 3. The connections among attributes proposed in models, standards and questionnaires are presented in Section 5 and 6.1 are drawn based on their definitions (Figure 7).

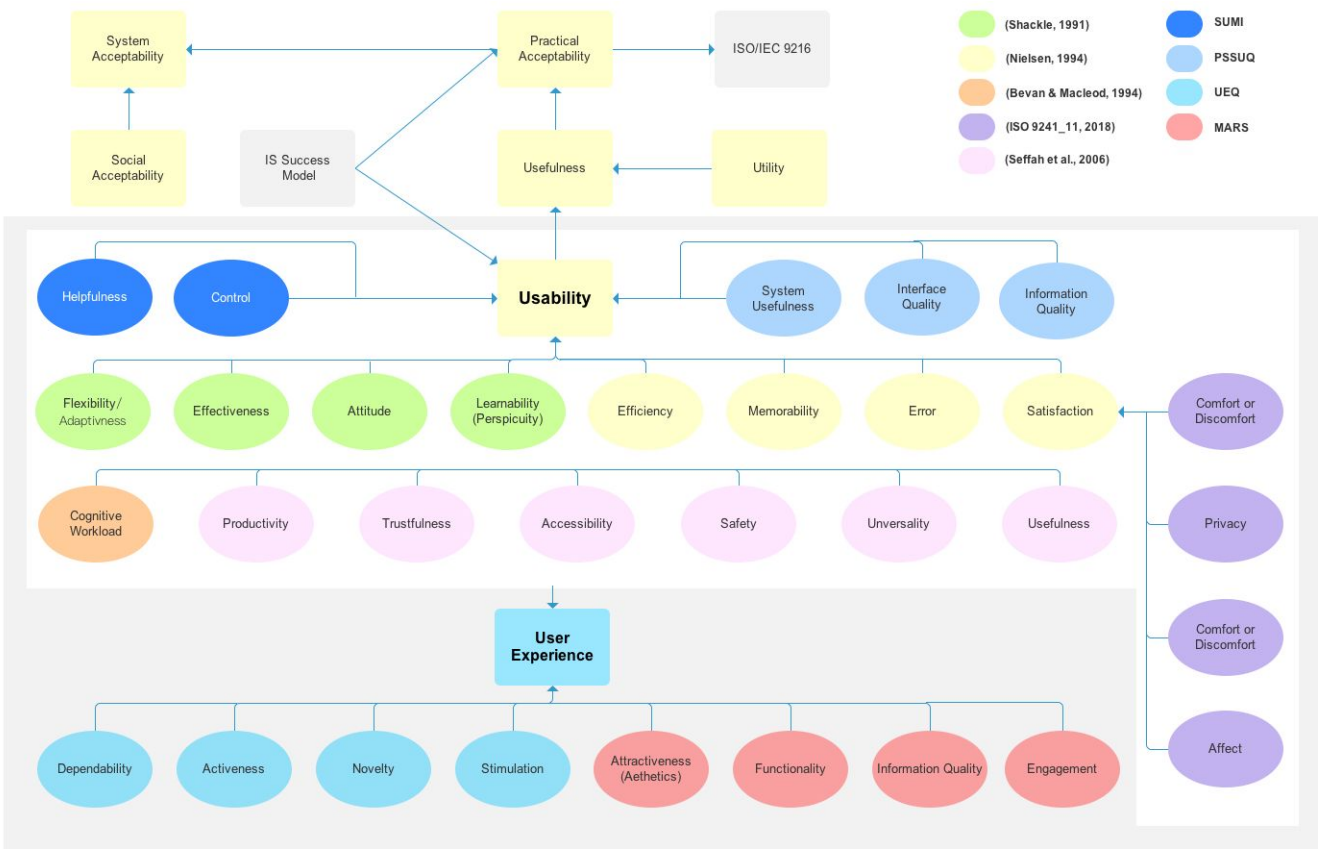


Figure 7: All attributes used in models, standards and questionnaires.

Even though a large corpus of literature has theorized usability and put efforts to clarify its definitions and specify the attributes and measures, the majority only lists all possible attributes that contribute to the usability and has not investigated the possible inner relations among these attributes. Also, the weight of satisfaction to usability has rarely been investigated. It would be convenient if each of the attributes of usability was equally important in every product and for every user, however they are not (Quesenbery, 2004).

Yet, it's plausible to analyze attributes' internal relationships and each one's weights to usability. A usability study that assessed the usability of Duolingo's desktop website and mobile phone applications also investigated the relationships between the four independent variables (learnability, efficiency, few errors, and attractiveness) and one dependent variable (satisfaction) (Nurhudatiana et al., 2018). They found that efficiency and errors had a significant influence on user satisfaction in desktop website usage, while only attractiveness showed a significant influence on user satisfaction in mobile app usage.

Meanwhile, satisfaction's contribution to usability has been evaluated. Both of the IS success model and mobile catering application success model consider satisfaction as a significant reason and indicator for the success of a system. The weight of satisfaction has been given theoretically and empirically and validated by data. Moreover, the self - determination theory and two-factor theory rigorously elaborate on life satisfaction and job satisfaction respectively, which can show that satisfaction has been placed in an important position in other fields. Despite the two groups of terms (success and usability; life satisfaction, job satisfaction, and user satisfaction) that do not have the exact same scope, these models and theories give some insights to reflecting upon the importance of user satisfaction to usability.

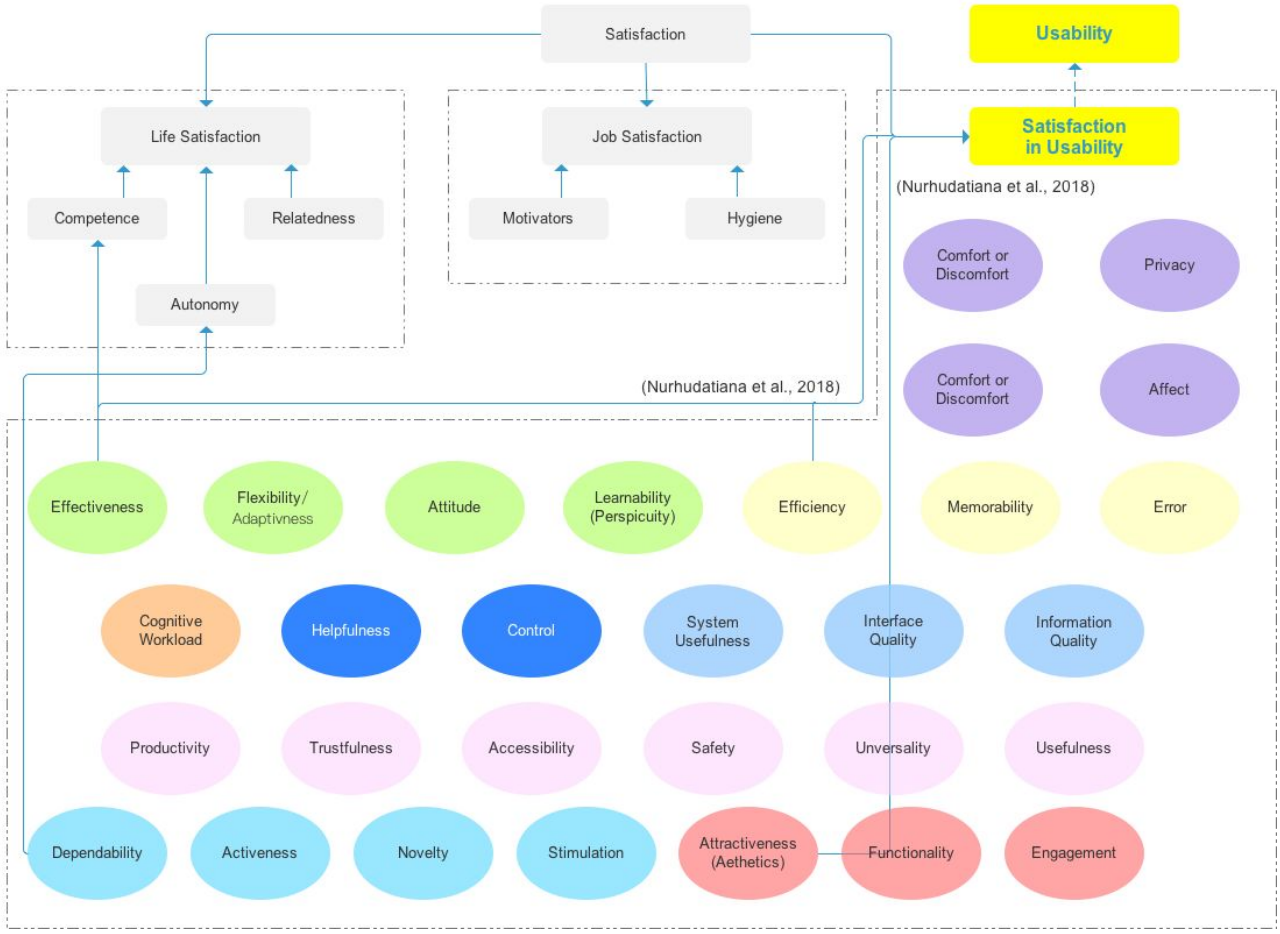


Figure 8: Satisfaction and Usability

Here we present a hypothesis (Figure 8) that satisfaction increases as the improvement of effectiveness, efficiency, functionality, information quality, etc.while satisfaction decreases as cognitive workload accumulates.

This hypothesis emphasizes cognitive workload, especially in mobile phone application's usability. The cognitive workload is initially mentioned as one independent attribute in the MUSiC project back in the 1980s, which has been neglected or merged into efficiency. In ISO 9421_11 (2018), efficiency is the resources used in relation to the results achieved and the resources include time, human effort, money, and materials. The human effort used is the mental and physical effort expended to complete specified tasks. The mental effort can be viewed as a cognitive workload. Two unique features of mobile phones mentioned in Section 5 are the reason why cognitive workload becomes a separate attribute of usability.

9. Conclusion and Future Direction

This paper reviews usability attributes and measures in usability testing, as well as how usability testing being adopted to mobile phone applications. The variance of satisfaction's definitions and measurements drew our attention to further investigate. After reviewing a relevant set of satisfaction theories and models in other fields, we proposed a usability model with a focus on satisfaction. Yet, it still lacks more literature research for theoretical support and data collection for validation. In the future, we will further investigate and validate the

attributes of usability via two approaches used in Duolingo's usability study (Nurhudatiana et al., 2018) and the mobile catering application success model (Wang et al., 2019).

Reference

Adhy, S., Prasetio, A., Noranita, B., & Saputra, R. (2018). Usability Testing of Weather Monitoring on Android Application. *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, 1–6. https://doi.org/10.1109/ICICOS.2018.8621752

Ahmed, S., Abubakar, A., Ibrahim, H., Garry, T., & Andrew, T. (2018). YORwalK: Desiging a Smartphone Exercise Application for People with Intermittent Claudication. *Studies in Health Technology and Informatics*, 311–315. https://doi.org/10.3233/978-1-61499-852-5-311

Alkhafaji, A., Cocea, M., Crellin, J., & Fallahkhair, S. (2017). Guidelines for designing a smart and ubiquitous learning environment with respect to cultural heritage. *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, 334–339. https://doi.org/10.1109/RCIS.2017.7956556

Alreck, P. L., & Settle, R. B. (1985). *The Survey Research Handbook. Homewood, IL: Richard D. Irwin*. Inc.

Alshamari, M., & Mayhew, P. (2009). Technical Review: Current Issues of Usability Testing. *IETE Technical Review*, *26*(6), 402–406. https://doi.org/10.4103/0256-4602.57825

Alturki, R., & Gay, V. (2019). Usability Attributes for Mobile Applications: A Systematic Review. In M. A. Jan, F. Khan, & M. Alam (Eds.), *Recent Trends and Advances in Wireless and IoT-enabled Networks* (pp. 53–62). https://doi.org/10.1007/978-3-319-99966-1_5

Alva, M. E. O., Martínez P., A. B., Cueva L., J. M., Sagástegui Ch., T. H., & López P., B. (2003). Comparison of Methods and Existing Tools for the Measurement of Usability in the Web. *Web Engineering*, 386–389. Springer Berlin Heidelberg.

Aragon, M. C., Castillo, R., Agustin, J., & Aguilar, I. B. (2019). Utilization of Feature Detector Algorithms in a Mobile Signature Detector Application. *Proceedings of the 2019 2nd International Conference on Information Science and Systems - ICISS 2019*, 49–53. https://doi.org/10.1145/3322645.3322701

Aykin, N. M., & Aykin, T. (1991). Individual differences in human-computer interaction. *Computers & Industrial Engineering*, *20*(3), 373–379. https://doi.org/10.1016/0360-8352(91)90009-U

Bai, B., Law, R., & Wen, I. (2008). The impact of website quality on customer satisfaction and purchase intentions: Evidence from Chinese online visitors. *International Journal of Hospitality Management*, *27*(3), 391–402. https://doi.org/10.1016/j.ijhm.2007.10.008

Barnum, C. M. (2010). *Usability testing essentials: Ready, set... Test!* Elsevier.

Baskoro, K., & Widyanti, A. (2018). Usability Evaluation on an Indonesian Mobile Application for Small Business Lending. *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*, 148–153. https://doi.org/10.1109/ICITSI.2018.8695957

Bastien, J. M. C. (2010). Usability testing: A review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, *79*(4), e18–e23. https://doi.org/10.1016/j.ijmedinf.2008.12.004

Beatty, A. L., Magnusson, S. L., Fortney, J. C., Sayre, G. G., & Whooley, M. A. (2018). VA FitHeart, a Mobile App for Cardiac Rehabilitation: Usability Study. *JMIR Human Factors*, *5*(1), e3. https://doi.org/10.2196/humanfactors.8017

Bennett, J. L. (1979). The commercial impact of usability in interactive systems. *Man-Computer Communication, Infotech State-of-the-Art*, *2*, 1–17.

Bennett, J. L. (1984). Managing to meet usability requirements: Establishing and meeting software development goals. *Visual Display Terminals, Prentice-Hall*, 161–184.

BEVAN, N., & MACLEOD, M. (1994). Usability measurement in context. *Behaviour & Information Technology*, *13*(1–2), 132–145. https://doi.org/10.1080/01449299408914592

Birnstiel, S., Steinmüller, B., Bissinger, K., Doll-Gerstendörfer, S., & Huber, S. (2019). Gartenfreund: Exploring the Botanical Garden with an Inclusive App. *Proceedings of Mensch Und Computer 2019*, 499–502. https://doi.org/10.1145/3340764.3344446

Bitner, M. J., & Hubbert, A. R. (1994). Encounter Satisfaction versus Overall Satisfaction versus Quality: The Customer's Voice. In *Service Quality: New Directions in Theory and Practice* (pp. 72–94). https://doi.org/10.4135/9781452229102

Brassard, M. (1991). The Memory Jogger Plus+. *The Journal for Healthcare Quality (JHQ)*, *13*(5), 67.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brooke, J. (1996). *SUS - A quick and dirty usability scale*. 7.

Casida, J. M., Aikens, J. E., Craddock, H., Aldrich, M. W., & Pagani, F. D. (2018). Development and Feasibility of Self-Management Application in Left-Ventricular Assist Devices: *ASAIO Journal*, *64*(2), 159–167. https://doi.org/10.1097/MAT.0000000000000673

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an Instrument Measuring User Satisfaction of the Human-computer Interface. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 213–218. https://doi.org/10.1145/57167.57203

Choo, M. S., Jeong, S. J., Cho, S. Y., Yoo, C., Jeong, C. W., Ku, J. H., & Oh, S.-J. (2017). Development of Decision Support Formulas for the Prediction of Bladder Outlet Obstruction and Prostatic Surgery in Patients With Lower Urinary Tract Symptom/Benign Prostatic Hyperplasia: Part II, External

Validation and Usability Testing of a Smartphone App. *International Neurourology Journal*, *21*(Suppl 1), S66-75. https://doi.org/10.5213/inj.1734854.427

Constantine, L. L., & Lockwood, L. A. D. (1999). *Software for use a practical guide to the models and methods of usage-centered design*. Retrieved from http://proquest.safaribooksonline.com/9780768685305

Coursaris, C., & Kim, D. (2006). A qualitative review of empirical mobile usability studies. *AMCIS 2006 Proceedings*, 352.

Crosby, L. E., Ware, R. E., Goldstein, A., Walton, A., Joffe, N. E., Vogel, C., & Britto, M. T. (2017). Development and evaluation of iManage: A self-management app co-designed by adolescents with sickle cell disease: Crosby et al. *Pediatric Blood & Cancer*, *64*(1), 139–145. https://doi.org/10.1002/pbc.26177

Dabholkar, P. A., Shepherd, C. D., & Thorpe, D. I. (2000). A comprehensive framework for service quality: An investigation of critical conceptual and measurement issues through a longitudinal study. *Journal of Retailing*, *76*(2), 139–173. https://doi.org/10.1016/S0022-4359(00)00029-4

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*.

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

DeLone, W. H., & McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, *3*, 60–95. https://doi.org/10.1287/isre.3.1.60

Delone, W. H., & McLean, E. R. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *Journal of Management Information Systems*, *19*(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748

DeLone, W. H., & McLean, E. R. (2004). Measuring e-Commerce Success: Applying the DeLone & McLean Information Systems Success Model. *International Journal of Electronic Commerce*, *9*(1), 31–47. https://doi.org/10.1080/10864415.2004.11044317

Ding, E., Liu, D., Soni, A., Adaramola, O., Han, D., Bashar, S. K., … McManus, D. D. (2017). Impressions of Older Patients with Cardiovascular Diseases to Smart Devices for Heart Rhythm Monitoring. *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 270–271. https://doi.org/10.1109/CHASE.2017.97

Dumas, J. S., Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Intellect books.

Evans, D. C. (2017). *Bottlenecks: Aligning UX design with user psychology*. Apress.

Farrell, S. (2017). UX Research Cheat Sheet. Retrieved October 31, 2019, from Nielsen Norman Group website: https://www.nngroup.com/articles/ux-research-cheat-sheet/

Ferrer-Garcia, M., Pla-Sanjuanelo, J., Dakanalis, A., Vilalta-Abella, F., Riva, G., Fernandez-Aranda, F., … Gutiérrez-Maldonado, J. (2017). Eating behavior style predicts craving and anxiety experienced in food-related virtual environments by patients with eating disorders and healthy controls. *Appetite*, *117*, 284–293. https://doi.org/10.1016/j.appet.2017.07.007

Ferrer-Garcia, M., Pla-Sanjuanelo, J., Dakanalis, A., Vilalta-Abella, F., Riva, G., Fernandez-Aranda, F., … Gutiérrez-Maldonado, J. (2019). A Randomized Trial of Virtual Reality-Based Cue Exposure Second-Level Therapy and Cognitive Behavior Second-Level Therapy for Bulimia Nervosa and Binge-Eating Disorder: Outcome at Six-Month Followup. *Cyberpsychology, Behavior, and Social Networking*, *22*(1), 60–68. https://doi.org/10.1089/cyber.2017.0675

Garcia, A. C., & de Lara, S. M. A. (2018). Enabling Aid in Remote Care for Elderly People via Mobile Devices: The MobiCare Case Study. *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, 270–277. https://doi.org/10.1145/3218585.3218671

Garrett, J. J. (2010). *The Elements of User Experience: User-Centered Design for the Web and Beyond*. Pearson Education.

Goodhue, D. (1986). IS ATTITUDES: TOWARD THEORETICAL AND DEFINITION CLARITY. *ICIS 1986 Proceedings*. Retrieved from https://aisel.aisnet.org/icis1986/26

Gossain, S., & Anderson, B. (1990). An iterative-design model for reusable object-oriented software. *ACM SIGPLAN Notices*, *25*(10), 12–27.

Gould, J. D. (1988). How to design usable systems. In *Handbook of human-computer interaction* (pp. 757–789). Elsevier.

Gould, J. D., & Boies, S. J. (1983). Human factors challenges in creating a principal support office system—The speech filing system approach. *ACM Transactions on Information Systems (TOIS)*, *1*(4), 273–298.

Gould, J. D., Boies, S. J., Levy, S., Richards, J. T., & Schoonard, J. (1987). The 1984 Olympic Message System: A test of behavioral principles of system design. *Communications of the ACM*, *30*(9), 758–769.

Gould, J. D., & Lewis, C. (1985). Designing for Usability: Key Principles and What Designers Think. *Commun. ACM*, *28*(3), 300–311. https://doi.org/10.1145/3166.3170

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), 1. https://doi.org/10.1186/2194-0827-1-1

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (pp. 139–183). https://doi.org/10.1016/S0166-4115(08)62386-9

Hashim, A. S., & Lee, V. J. X. (2018). Usability of ShopCart Among Customers at Shopping Malls. *2018 IEEE Conference on E-Learning, e-Management and e-Services (IC3e)*, 140–144. https://doi.org/10.1109/IC3e.2018.8632655

Hernandez, N., Cruciani, F., Favela, J., McChesney, I., Zhang, S., Nugent, C., & Cleland, I. (2019). Preliminary evaluation of a self-management health app by people with cognitive impairment. *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 535–540. https://doi.org/10.1109/PERCOMW.2019.8730755

Herzberg, F. (1974). *The motivation-hygiene concept and problems of manpower*. BYU Electronic Media Department.

Hoober, S., & Berkman, E. (2011). *Designing Mobile Interfaces: Patterns for Interaction Design*. O'Reilly Media, Inc.

Hughes, C., Mariscal, T., Baye, M., Belay, G. J., Hintze, A., Padilla, A., … Gordon-Murer, C. (2019). Development of an Upper Extremity {Stroke Rehabilitation mHealth Application for sub-Saharan Africa: A Usability Study. *2019 IST-Africa Week Conference (IST-Africa)*, 1–8. https://doi.org/10.23919/ISTAFRICA.2019.8764867

Igersheim, R. H. (1976). Managerial Response to an Information System. *Proceedings of the June 7-10, 1976, National Computer Conference and Exposition*, 877–882. https://doi.org/10.1145/1499799.1499918

Johnson, J. (2013). *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines*. Elsevier.

Johnson, J., & Henderson, A. (2011). *Conceptual Models: Core to Good Design*. Morgan & Claypool Publishers.

Jones, M. A., & Suh, J. (2000). Transaction-specific satisfaction and overall satisfaction: An empirical analysis. *Journal of Services Marketing*, *14*(2), 147–159. https://doi.org/10.1108/08876040010371555

Kaptelinin, V., & Nardi, B. A. (2006). *Acting with Technology: Activity Theory and Interaction Design*. MIT Press.

Kascak, L., Rébola, C. B., Braunstein, R., & Sanford, J. (2013). Mobile Application Concept Development for Remote Patient Monitoring. *2013 IEEE International Conference on Healthcare Informatics*, 545–550. https://doi.org/10.1109/ICHI.2013.85

Khan, M. N. R., Sonet, H. H., Yasmin, F., Yesmin, S., Sarker, F., & Mamun, K. A. (2017). 'Bolte Chai'—An Android application for verbally challenged children. *2017 4th International Conference on Advances in Electrical Engineering (ICAEE)*, 541–545. https://doi.org/10.1109/ICAEE.2017.8255415

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*(3), 210–212. https://doi.org/10.1111/j.1467-8535.1993.tb00076.x

Kumar, A., Srivastava, K., Yadav, K., & Deshmukh, O. (2017). Multi-faceted Index Driven Navigation for Educational Videos in Mobile Phones. *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, 357–361. https://doi.org/10.1145/3025171.3025221

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger (Ed.), *HCI and Usability for Education and Work* (pp. 63–76). Springer Berlin Heidelberg.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 719–728. ACM.

Levin, M. (2014). *Designing Multi-Device Experiences: An Ecosystem Approach to User Experiences across Devices*. O'Reilly Media, Inc.

Lewis, J. R. (1992). Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *36*(16), 1259–1260. https://doi.org/10.1177/154193129203601617

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, *7*(1), 57–78. https://doi.org/10.1080/10447319509526110

Lewis, J. R. (2002). Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human–Computer Interaction*, *14*(3–4), 463–488. https://doi.org/10.1080/10447318.2002.9669130

Lewis, J. R. (2006). Usability testing. *Handbook of Human Factors and Ergonomics*, *12*, e30.

Lucas, H. C. (1978). Empirical Evidence for a Descriptive Model of Implementation. *MIS Quarterly*, *2*(2), 27–42. https://doi.org/10.2307/248939

Lund, A. M. (2001). Measuring usability with the use questionnaire12. *Usability Interface*, *8*(2), 3–6.

Mallan, V. S., Gopi, S., Muir, A., & Bhavani, R. R. (2017). Comparative empirical usability assessment of two HRI input devices for a mobile robot. *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*, 331–337. https://doi.org/10.1109/ISPCC.2017.8269699

Martin, P. Y., & Turner, B. A. (1986). Grounded theory and organizational research. *The Journal of Applied Behavioral Science*, *22*(2), 141–157.

McKinney, V., Yoon, K., & Zahedi, F. "Mariam." (2002). The Measurement of Web-Customer Satisfaction: An Expectation and Disconfirmation Approach. *Information Systems Research*, *13*(3), 296–315. https://doi.org/10.1287/isre.13.3.296.76

Mendoza, A. (2013). *Mobile User Experience: Patterns to Make Sense of it All*. Newnes.

Miller, R. B. (1971). *Human ease of use criteria and their tradeoffs*. IBM, Systems Development Division, Poughkeepsie Lab.

Navarro, C. X., Molina, A. I., & Redondo, M. A. (2015). Towards a Model for Evaluating the Usability of M-learning Systems: From a Mapping Study to an Approach. *IEEE Latin America Transactions*, *13*(2), 552–559. https://doi.org/10.1109/TLA.2015.7055578

Nayebi, F., Desharnais, J.-M., & Abran, A. (2012). The state of the art of mobile application usability evaluation. *2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1–4. https://doi.org/10.1109/CCECE.2012.6334930

Nazar, M., & Zulfadli, Z. (2017). Usability testing of chemistry dictionary (ChemDic) developed on Android studio. *2017 International Conference on Electrical Engineering and Informatics (ICELTICs)*, 221–225. https://doi.org/10.1109/ICELTICS.2017.8253265

Neil, T. (2014). *Mobile Design Pattern Gallery: UI Patterns for Smartphone Apps*. O'Reilly Media, Inc.

Newbery, P., & Farnham, K. (2013). *Experience Design: A Framework for Integrating Brand, Experience, and Value*. John Wiley & Sons.

Ng, C., Cheong, S., Hajimohammadhosseinmemar, E., & Yap, W. (2017). Mobile outdoor parking space detection application. *2017 IEEE 8th Control and System Graduate Research Colloquium (ICSGRC)*, 81–86. https://doi.org/10.1109/ICSGRC.2017.8070573

Nielsen, J. (1993). Iterative user-interface design. *Computer*, *26*(11), 32–41.

Nielsen, J. (1994). *Usability engineering*. Elsevier.

Nielsen, J., & Budiu, R. (2013). *Mobile Usability*. MITP-Verlags GmbH & Co. KG.

Nielsen, J., & Norman, D. (2014). The definition of user experience. *Nielsen Norman Group*, 191.

Norman, D. A. (1988). *The psychology of everyday things*. New York, NY, US: Basic Books.

Norman, D. A. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books.

Norman, D., Miller, J., & Henderson, A. (1995). What you see, some of what's in the future, and how we go about doing it: HI at Apple Computer. *Conference Companion on Human Factors in Computing Systems*, 155. ACM.

Nugraha, A. P., Syaifullah, D. H., & Puspasari, M. A. (2018). Usability Evaluation of Main Function on Three Mobile Banking Application. *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, *3*, 1–6. https://doi.org/10.1109/ICIIBMS.2018.8549998

Nurhudatiana, A., Hiu, A. N., & Ce, W. (2018). Should I Use Laptop or Smartphone? A Usability Study on an Online Learning Application. *2018 International Conference on Information Management and Technology (ICIMTech)*, 565–570. https://doi.org/10.1109/ICIMTech.2018.8528134

Oliva, T. A., Oliver, R. L., & MacMillan, I. C. (1992). A Catastrophe Model for Developing Service Satisfaction Strategies. *Journal of Marketing*, *56*(3), 83–95. https://doi.org/10.1177/002224299205600306

Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, *17*(4), 460–469. https://doi.org/10.1177/002224378001700405

Oliver, R. L. (1993). Cognitive, Affective, and Attribute Bases of the Satisfaction Response. *Journal of Consumer Research*, *20*(3), 418–430. https://doi.org/10.1086/209358

Paldán, K., Simanovski, J., Ullrich, G., Steinmetz, M., Rammos, C., Jánosi, R. A., … Lortz, J. (2019). Feasibility and Clinical Relevance of a Mobile Intervention Using TrackPAD to Support Supervised Exercise Therapy in Patients With Peripheral Arterial Disease: Study Protocol for a Randomized Controlled Pilot Trial. *JMIR Research Protocols*, *8*(6), e13651. https://doi.org/10.2196/13651

Paslakis, G., Fauck, V., Röder, K., Rauh, E., Rauh, M., & Erim, Y. (2017). Virtual reality jogging as a novel exposure paradigm for the acute urge to be physically active in patients with eating disorders: Implications for treatment. *International Journal of Eating Disorders*, *50*(11), 1243–1246. https://doi.org/10.1002/eat.22768

Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, *20*(1), 61. https://doi.org/10.1007/BF02723476

Pine, B. J., & Gilmore, J. H. (2011). *The Experience Economy*. Harvard Business Press.

Pla-Sanjuanelo, J., Ferrer-Garcia, M., Vilalta-Abella, F., Riva, G., Dakanalis, A., Ribas-Sabaté, J., … Gutierrez-Maldonado, J. (2017). VR-based cue-exposure therapy (VR-CET) versus VR-CET plus pharmacotherapy in the treatment of bulimic-type eating disorders. *Annual Review of CyberTherapy and Telemedicine*, *15*, 116–122. Retrieved from Scopus.

Pla-Sanjuanelo, J., Ferrer-García, M., Vilalta-Abella, F., Riva, G., Dakanalis, A., Ribas-Sabaté, J., … Gutiérrez-Maldonado, J. (2019). Testing virtual reality-based cue-exposure software: Which cue-elicited responses best discriminate between patients with eating disorders and healthy controls? *Eating and Weight Disorders*, *24*(4), 757–765. https://doi.org/10.1007/s40519-017-0419-4

Pratama, M., Setiawan, N. A., & Wibirama, S. (2017). User interface design for android-based family genealogy social media. *2017 7th International Annual Engineering Seminar (InAES)*, 1–5. https://doi.org/10.1109/INAES.2017.8068557

Preece, J. (2000). *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc.

Quesenbery, W. (2003). *Dimensions of Usability: Defining the Conversation, Driving the Process*. 8.

Quinn, C. C., Staub, S., Barr, E., & Gruber-Baldini, A. (2019). Mobile Support for Older Adults and Their Caregivers: Dyad Usability Study. *JMIR Aging*, *2*(1), e12276. https://doi.org/10.2196/12276

Reis, A., Coutinho, F., Ferreira, J., Tonelo, C., Ferreira, L., & Quintas, J. (2019). Monitoring System for Emergency Service in a Hospital Environment. *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, 1–4. https://doi.org/10.1109/ENBENG.2019.8692461

Rodriguez, D., Dehghan, M., Figueroa, P., & Jagersand, M. (2018). Evaluation of Smartphone-based Interfaces for Navigation Tasks in Unstructured Environments for Ground Robots. *2018 IEEE 2nd*

Colombian Conference on Robotics and Automation (CCRA), 1–6. https://doi.org/10.1109/CCRA.2018.8588148

Ryan, R. M., & Deci, E. L. (2008). From Ego Depletion to Vitality: Theory and Findings Concerning the Facilitation of Energy Available to the Self. *Social and Personality Psychology Compass*, *2*(2), 702–717. https://doi.org/10.1111/j.1751-9004.2008.00098.x

Saputra, K., Farhan, K., & Irvanizam, I. (2018). Analysis on the Comparison of Retrofit and Volley Libraries on Android-Based Mosque Application. *2018 International Conference on Electrical Engineering and Informatics (ICELTICs)*, 117–121. https://doi.org/10.1109/ICELTICS.2018.8548881

Scupin, R. (1997). The KJ method: A technique for analyzing data derived from Japanese ethnology. *Human Organization*, 233–237.

Seffah, A., Kececi, N., & Donyaee, M. (2001). QUIM: A framework for quantifying usability metrics in software quality models. *Proceedings Second Asia-Pacific Conference on Quality Software*, 311–318. https://doi.org/10.1109/APAQS.2001.990036

Shackel, B. (1984). The concept of usability. *Visual Display Terminals: Usability Issues and Health Concerns*, 45–87.

SHACKEL, B. (1991). USABILITY—CONTEXT, FRAMEWORK, DEFINITION, DESIGN AND EVALUATION. *Human Factors for Informatics Usability*, 21.

Shada, G. S., & Ayu, M. A. (2018). Designing Android User Interface for University Mobile Library. *2018 International Conference on Computing, Engineering, and Design (ICCED)*, 224–229. https://doi.org/10.1109/ICCED.2018.00051

Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. (2015). Mobile App Rating Scale: A New Tool for Assessing the Quality of Health Mobile Apps. *JMIR MHealth and UHealth*, *3*(1). https://doi.org/10.2196/mhealth.3422

Sumanasekera, K., Mihilar, S., Wickramasinghe, C., & Arunathilake, S. (2018). "Kawulu": A Voice based Social Network using Smart Mobile Devices for the Visually Impaired Community in Sri Lanka. *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 228–233. https://doi.org/10.1109/ICTC.2018.8539569

Szymanski, D. M., & Hise, R. T. (2000). E-satisfaction: An initial examination. *Journal of Retailing*, *76*(3), 309–322. https://doi.org/10.1016/S0022-4359(00)00035-X

Tomaschko, M., & Hohenwarter, M. (2018). Usability Evaluation of a Mobile Graphing Calculator Application Using Eye Tracking. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies. Design, Development and Technological Innovation* (Vol. 10924, pp. 180–190). https://doi.org/10.1007/978-3-319-91743-6_14

Unrau, R., & Kray, C. (2019). Usability evaluation for geographic information systems: A systematic literature review. *International Journal of Geographical Information Science*, *33*(4), 645–665. https://doi.org/10.1080/13658816.2018.1554813

Veale, G., Dogan, H., & Murphy, J. (2019). Development and Usability Evaluation of a Nutrition and Lifestyle Guidance Application for People Living with and Beyond Cancer. In A. Marcus & W. Wang (Eds.), *Design, User Experience, and Usability. Application Domains* (Vol. 11585, pp. 337–347). https://doi.org/10.1007/978-3-030-23538-3_26

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *Proceedings of the Human Factors Society Annual Meeting*, *34*, 291–294. SAGE Publications Sage CA: Los Angeles, CA.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, *34*(4), 457–468.

Walker, J. L. (1995). Service encounter satisfaction: Conceptualized. *Journal of Services Marketing*. https://doi.org/10.1108/08876049510079844

Wang, Y.-S., Tseng, T. H., Wang, W.-T., Shih, Y.-W., & Chan, P.-Y. (2019). Developing and validating a mobile catering app success model. *International Journal of Hospitality Management*, *77*, 19–30. https://doi.org/10.1016/j.ijhm.2018.06.002

Wardhana, S., Sabariah, M. K., Effendy, V., & Kusumo, D. S. (2017). User interface design model for parental control application on mobile smartphone using user centered design method. *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, 1–6. https://doi.org/10.1109/ICoICT.2017.8074715

Westbrook, R. A. (1980). A Rating Scale for Measuring Product/ Service Satisfaction. *Journal of Marketing*, *44*(4), 68–72. https://doi.org/10.1177/002224298004400410

Westbrook, R. A., & Oliver, R. L. (1991). The Dimensionality of Consumption Emotion Patterns and Consumer Satisfaction. *Journal of Consumer Research*, *18*(1), 84–91. https://doi.org/10.1086/209243

Wichiennit, N., Sunat, K., Chiewchanwattana, S., Louchaisa, B., & Onnoom, B. (2017). Design and development of application for crime scene notification system. *2017 10th International Conference on Ubi-Media Computing and Workshops (Ubi-Media)*, 1–6. https://doi.org/10.1109/UMEDIA.2017.8074103

Wohlfahrt-Laymann, J., Hermens, H., Villalonga, C., Vollenbroek-Hutten, M., & Banos, O. (2018). Enabling remote assessment of cognitive behaviour through mobile experience sampling. *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 794–799. https://doi.org/10.1109/PERCOMW.2018.8480310

Wolfinbarger, M., & Gilly, M. C. (2003). eTailQ: Dimensionalizing, measuring and predicting etail quality. *Journal of Retailing*, *79*(3), 183–198. https://doi.org/10.1016/S0022-4359(03)00034-4

Woods, L. S., Duff, J., Roehrer, E., Walker, K., & Cummings, E. (2019). Patients' Experiences of Using a Consumer mHealth App for Self-Management of Heart Failure: Mixed-Methods Study. *JMIR Human Factors*, *6*(2), e13009. https://doi.org/10.2196/13009

Yabut, E. R., Balceda, C. D., Juan, R. E. Q. S., Tumamak, J. R., Velasquez, R. E., Jamis, M. N., & Manuel, R. E. (2017). e-wasBaha: A mobile application framework for flood monitoring in Metro Manila using crowdsourcing. *2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information*

Technology, Communication and Control, Environment and Management (HNICEM), 1–6. https://doi.org/10.1109/HNICEM.2017.8269533

Yuksel, A., & Yuksel, F. (2001). Measurement and Management Issues in Customer Satisfaction Research: Review, Critique and Research Agenda: Part Two. *Journal of Travel & Tourism Marketing*, *10*(4), 81–111. https://doi.org/10.1300/J073v10n04_04

Zaror, C., Espinoza-Espinoza, G., Atala-Acevedo, C., Muñoz-Millán, P., Li, Y., Clarke, K., … Mariño, R. (2019). Validation and usability of a mobile phone application for epidemiological surveillance of traumatic dental injuries. *Dental Traumatology*, *35*(1), 33–40. https://doi.org/10.1111/edt.12444

Zhang, D., & Adipat, B. (2005). Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications. *International Journal of Human–Computer Interaction*, *18*(3), 293–308. https://doi.org/10.1207/s15327590ijhc1803_3