A Tertiary Study on Quality in Use Evaluation of Smart Environment Applications

Maria Paula Corrêa Angeloni[1][0000−0002−1129−5073], Rafael Duque Medina[2][0000−0001−8636−3213], Káthia Marçal de Oliveira[1][0000−0001−8146−5966], Emmanuelle Grislin-Le Strugeon[1,3][0000−0002−8429−4012], and Cristina Tirnauca[2][0000−0002−7129−2237]

[1] UPHF, CNRS, UMR 8201 - LAMIH, Valenciennes, France
[2] Depart. de Matemáticas, Estadística y Computación, Universidad de Cantabria, Santander, Spain
[3] INSA Hauts-de-France, Valenciennes, France

Abstract. As the population grows older, the need for special assistance increases, and a modern alternative to mitigate the absence of face-to-face caregivers (which is expensive) is to take advantage of technological devices in so called smart environments, which can be an economical and practical solution. Guaranteeing the software quality of applications in these spaces before providing it to end users is essential, especially in situations involving senior citizens or people with motor disabilities. In order to investigate how the quality evaluation of smart environment applications has been performed, we carried out a tertiary study. From a total of 1,028 studies, 21 were carefully selected for analysis. The results confirmed that classical questionnaires and interviews are the techniques that are still used the most for evaluation, but that simulation appears as a new trend to that end.

Keywords: Quality in Use · Smart environments · Literature review.

1 Introduction

Smart environments have been considered as an appropriate living alternative for both ageing safely at home and also receiving care as needed [16]. A smart home can be described as a "residence wired with technology features that monitor the well-being and activities of their residents to improve overall quality of life, increase independence and prevent emergencies" [13]. Ensuring the quality of these systems is essential if they are to be effectively utilized: without a solid reference on user evaluation, a system evaluation cannot be tackled well [30].

The quality of a system application is defined as "the degree to which the system satisfies the stated and implied needs of its various stakeholders and thus provides value" [26]. These needs are represented in the SQuaRE International Standards by two quality models: (i) *a product quality model* made up of characteristics (such as functional suitability, performance efficiency and maintainability) related to the static properties of the software and the dynamic

properties of the computer system; and (ii) a *Quality-in-Use (QinU) model* made up of characteristics (such as effectiveness and efficiency) related to the outcome of the human-computer interaction when a product is used in a specific context of use. While the first model assesses the quality of the product itself, the QinU perspective assesses the effect of the interaction between the user and the software, taking into consideration the user experience (UX). We focused this study on QinU, as our main intention is to evaluate the quality of software used in smart environment contexts.

Given our main interest in evaluating QinU for software applications in smart homes that are usually inhabited by elderly people, we wondered how these evaluations have been carried out and whether the classic user evaluation sessions have been applied. The QinU evaluation of smart environment applications involves placing users in the smart environment (in this case, smart homes) for evaluation sessions due to the array of sensors that usually capture information for this type of application. Evaluations like that may be difficult or even unsafe for the elderly and/or people with reduced mobility; therefore, we should secure the QinU even before carrying them out. Aiming to find out how researchers are addressing this problem and which approaches have been applied for evaluating smart home applications that could address this situation properly, we decided to carry out a tertiary study [12] by rapid review [45].

The rest of this paper is organized as follows: Section 2 briefly presents basic concepts of QinU. Section 3 describes the research protocol and execution procedures of the study. Then, Section 4 discusses the results of the performed review and Section 5 the threats to the validity of this study. Finally, Section 6 presents some final remarks and our ongoing work.

2 Background

QinU considers how much a software can address the user needs in a specific context of use and it is divided into five categories [26]: *effectiveness*, for how accurately users can perform their intended tasks; *efficiency*, for how easily and fast such tasks can be accomplished; *satisfaction*, for how pleased the users are with the product; *freedom from risk*, for how much such product lessens potential risks related to either the environment, economic status or humans' health; and *context coverage*, for the degree to which the product can be used with the four previous characteristics in specified contexts of use (context completeness) and also in contexts beyond those initially explicitly identified (flexibility).

When we talk about evaluating interactive systems, we immediately turn to usability issues, as presented in the ISO 9241-11 standard [21]. Usability is defined as "the degree to which a product can be used, by identified users, to achieve defined objectives with effectiveness, efficiency and satisfaction, in a specified context of use". Effectiveness, efficiency and satisfaction, which are the tripod of the ISO/IEC 9241-11 definition, are present in the QinU model of the SQUaRE standard [26].

The evaluation of QinU and usability issues has been largely explored in literature with methods and applications. We can quote, for instance: mathematical simulations (e.g. [10]), where the authors establish values and apply them to mathematical formulas for validation, or agent-based simulations (e.g. [8]), to emulate humans' behaviour. One of the most common ways of evaluating QinU and usability is through questionnaires that can be answered by end users (e.g. [38]) or by domain experts (e.g. [40]). In such situations, users must either interact with the evaluated application or have it demonstrated to them before answering a set of questions. Some authors decide to apply ad-hoc questionnaires, that is, a group of questions created for assessing a specific application with no intentions of recreating the process, or published questionnaires that are already established with a pattern of interrogations. Among the latter, there may be standardized questionnaires [5] or non-standardized ones.

3 The Tertiary Study

A tertiary study is described as a "systematic review of systematic reviews" that can be performed in a field where a number of reviews related to the same research questions has already been made [28]. In this section, we describe the planning (Section 3.1) and execution (Section 3.2) of our tertiary study on QinU of smart environment applications[4].

3.1 Planning: Research Protocol

From the main issues presented previously, the research questions (RQs) were defined to guide the information that should be extracted from the studies:

- RQ1 (main RQ): What are the most common evaluation approaches for QinU in smart environments?
- RQ2: What are the most evaluated types of systems regarding smart environments?
- RQ3: Which quality characteristics were the most evaluated in smart environments?

To find the secondary studies, the databases Scopus and Web of Science were selected considering that they are the most widely used databases for analysis [44]. Besides, they are largely used for systematic studies and gather most of the Computer Science references including ACM, Elsevier, IEEE and Springer.

The search string to be ran in such databases was defined using the PICOC strategy [42]: *P*opulation, for all studies related to QinU[5], considering all its subcharacteristics; *I*ntervention, for finding evaluations and assessments for the quality previously mentioned; *C*omparison, which was not included since we

[4] The detailed process and replication packages are available on [1, 2].

[5] The acronym QinU was not included, as we tried it on its own in the search string and no studies came up.

did not know of any other reviews with the same goal; *O*utcome, to establish that the results should include secondary studies; *C*ontext, to include intelligent environments, pervasive systems, and so on.

The search string was defined as follows, and each part was then linked with an "and" connector:

- **Population**: ("usability" OR "quality in use" OR "effectiveness" OR "efficiency" OR "satisfaction" OR "freedom from risk" OR "context*" OR "usefulness" OR "trust" OR "pleasure" OR "comfort" OR "flexibility")
- **Intervention**: ("evaluation" OR "assessment" OR "quality evaluation" OR "quality assessment")
- **Outcome**: ("*systematic literature review" OR "systematic* review*" OR "mapping study" OR "systematic mapping" OR "structured review" OR "secondary study" OR "literature survey" OR "review of survey*" OR "state of research" OR "state of art" OR "rapid review" OR "SLR" OR "scoping review")
- **Context**: ("smart*" OR "intelligent environment" OR "pervasive" OR "ubiquitous" OR "AAL" OR "ambient intelligence" OR "assisted living" OR "systems of systems" OR "internet of things" OR "Cyber-Physical Systems" OR "Industry 4" OR "fourth industrial revolution" OR "web of things" OR "Internet of Everything" OR "IoT" OR "CPS")

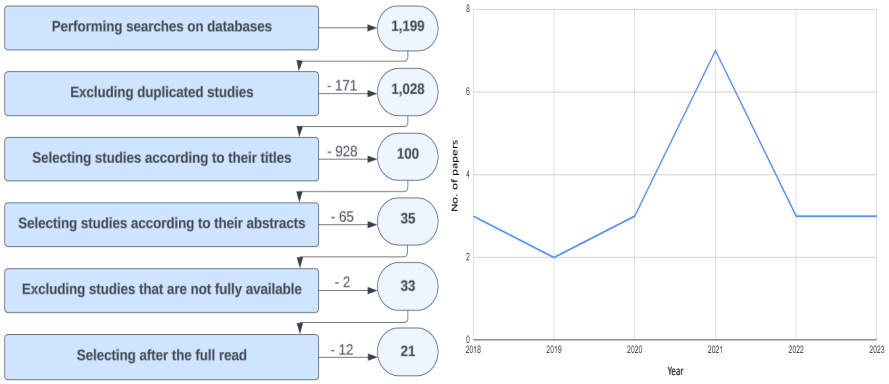
To be included in this tertiary study, each one of the reviewed studies had to comply with the following aspects of the defined inclusion criteria (IC):

- IC1: Published in the Computer Science or Engineering area;
- IC2: Published in the proceedings of a conference, journal, or as a book chapter;
- IC3: Published as a secondary study;
- IC4: Studies concerning the evaluation of the QinU of smart environment applications.

The exclusion criteria (EC) was established as well:

- EC1: Studies not published in English;
- EC2: Editorials, books or erratums;
- EC3: Studies duplicated in the results;
- EC4: Studies that have not been found fully available.

For extracting data, it was established that the following information would be collected from each one of the studies: (i) goal; (ii) number of papers reviewed; (iii) approaches used for evaluation; (iv) type of the system that was evaluated; (v) quality characteristics that were evaluated; (vi) type of secondary study; (vii) type of publication; and (viii) year of publication. A spreadsheet file and a Google form (that stores the data in a spreadsheet) were used, respectively, for the review process and extraction of data. Both spreadsheets were then accessed by all reviewers via Google Sheets.



(a) Research steps and their results. (b) Studies published per year.

Fig. 1: Research protocol and year distribution in the final selected set.

3.2 Execution

One PhD student and four professors expert on HCI and IoT applications participated in this phase. The primary selection was performed by one of the reviewers and peer-reviewed by another. In case of disagreements between these two, a third reviewer provides an opinion regarding the respective study to resolve the tie. In case of persistent doubt, a discussion with another reviewer was done to make a decision.

All selection criteria (IC, EC) was applied from the beginning of the research steps (Figure 1a). The search string was executed in both selected databases on July 19th, 2023. At this moment, thanks to the engine tools of the databases, IC1, IC2, EC1, and EC2 were already included in the search string automatically filtering the results: 338 from Scopus and 861 from Web of Science. Combining both results (1,199 papers), we could therefore apply EC3 to exclude duplicated studies before the manual review, which resulted in a total of 1028 studies. Then, all the titles were reviewed by one reviewer to include only secondary studies that worked to measure the QinU of software (IC3, IC4), which resulted in 100 studies. Then, all the abstracts were analyzed for the same purpose, which resulted in 35 studies. The next step implies downloading all papers for data extraction. However, 2 papers were not available (EC4) which resulted in 33 papers. After reading the full publications, we concluded that 12 articles concerned mobile applications and were not effectively applied in smart environments (IC4), which was not clear from the abstract, so they were also excluded. Finally, 21 studies remained for data extraction.

4 Discussion of the Results

We did not establish any date limits. We noticed that the studies brought in the search are very recent, ranging from 2018 to 2023, as shown in Figure 1b, with

at least two articles published each year and with its peak in 2021, which shows that not only it is a relatively new field of research, but also that it is a topic on which researchers are currently actively working on.

The secondary studies selected can be categorized based on the methodology used to choose the studies for analysis. In total, there were three different categories of secondary studies: *Systematic Literature Review* [4, 6, 14, 15, 18–20, 29, 33], *Scoping Review* [7, 9, 11, 31, 32, 34, 36, 37, 46], and *Systematic Mapping* [17, 39, 43].

Considering all the secondary studies analyzed, seven out of 21 (33,33%) were published in international conferences and 14 (66,66%) were published in journals. Not all the secondary studies listed the articles (primary studies) they analyzed.

4.1 RQ1. What are the most common evaluation approaches for QinU?

Different types of evaluation were found by the secondary studies (Table 1), with the most common evaluation approach being the questionnaire, whether the authors created it themselves for their study [4, 6, 7, 9, 15, 17, 20, 29, 31, 33, 43, 46] or applied an already published questionnaire [6, 7, 14, 17, 29, 31, 33, 34, 37, 39, 46], followed by interviews [4,6,7,9,17,20,29,31,33,34,36,37,43,46], focus groups [9,17, 20,29,31,33,34,46], observation [4,7,17,18,20,29,34,46], log data [11,17,33,34,46], the think aloud protocol [7, 17, 29, 31], surveys [18, 32, 46], user feedback [9, 29], heuristic evaluations [29, 31], and others. Table 1 presents the reference for each secondary study, the number of primary studies they analyzed (#), and the methods and techniques identified (e.g. [4] analyzed 24 primary studies).

More than half of the secondary studies analyzed (12) reported the use of questionnaires that were created by the authors with the sole purpose of answering their research questions and without any plan for repetition [4, 6, 7, 9, 15, 17, 20, 29, 31, 33, 43, 46], while other authors used established questionnaire forms to reach their goals. Between those, the secondary studies indicated a total of 44 different questionnaires, including well-known and standardized ones, according to [5] and [14], that are presented in Table 2 (which shows not only the standardized questionnaires' names, but also how many secondary studies found them and how many primary studies were found applying such questionnaires in each secondary study). The most common used questionnaire was System Usability Scale (SUS), which was found by 9 secondary studies applied in its original form [4, 6, 7, 15, 17, 29, 31, 37, 46] and was also in an adapted way in three secondary studies [7, 17, 46]: one of them adapted the questionnaire according to the context [17] and the others [7, 46] customized it by putting it together with other established questionnaires. Between the latter, one of them [46] found it adapted with the NASA Task Load Index (NASA-TLX), while another secondary study [7] spotted SUS adapted in two different primary studies: one combined it with PSSUQ (Post-Study System Usability Questionnaire) and another combined it with CSUQ (Computer System Usability Questionnaire). Three different secondary studies [4, 17, 37] also found the application of the NASA-TLX

Table 1: Methods and techniques found in the primary studies (#).

Ref.	#	Ad-hoc quest.	Estab. quest.	Interview	Focus group	Observation	Log data	Think aloud	Survey	User feed.	Simulation	Heuristics	Auto. test	Real testbed	Prototype	Exp. protocol	Others
[4]	24	13	-	2	-	7	-	-	-	-	-	-	-	-	-	-	-
[6]	44	16	19	13	-	-	-	-	-	-	-	-	-	-	-	-	-
[7]	35	12	26	6	-	4	-	2	-	-	-	-	-	-	-	-	-
[9]	21	8	-	8	3	-	-	-	-	2	-	-	-	-	-	-	1
[11]	63	-	-	-	-	-	7	-	-	-	-	-	-	-	-	-	56
[14]	553	-	553	-	-	-	-	-	-	-	-	-	-	-	-	-	-
[15]	21	21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
[17]	65	29	41	8	1	2	17	8	-	-	-	-	-	-	-	-	-
[18]	15	-	-	-	-	10	-	-	9	-	-	-	-	-	-	-	-
[19]	146	-	-	-	-	-	-	-	-	-	76	-	-	37	11	-	22
[20]	29	1	-	6	4	1	-	-	-	-	-	-	-	-	-	-	17
[29]	22	2	2	2	2	1	-	1	-	2	-	11	5	-	-	-	-
[31]	133	39	66	37	13	-	-	45	-	-	18	-	-	-	-	-	57
[32]	9	-	-	-	-	-	-	-	6	-	-	-	-	-	-	3	1
[33]	34	2	2	1	1	-	9	-	-	-	-	-	-	-	-	19	-
[34]	111	-	9	4	2	1	101	-	-	-	-	-	-	-	-	-	-
[36]	8	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	4
[37]	51	-	29	24	-	-	-	-	-	-	-	-	-	-	-	-	-
[39]	12	-	10	-	-	-	-	-	-	-	-	-	-	-	-	2	-
[43]	10	7	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-
[46]	31	14	17	16	2	4	3	-	3	-	-	-	-	-	-	-	-

Table 2: Standardized questionnaires found in the secondary studies.

Questionnaire	**#Sec. studies**	**#Pri. studies per sec. studies**
System Usability Scale (SUS)	9	1 [4], 16 [6], 16 [7], 1 [15], 11 [17], 2 [29], X [37], 44 [31], 3 [46]
System Usability Scale (SUS) adapted	3	2 [7], 1 [17], 1 [46]
Technology Acceptance Model (TAM)	4	4 [7], 5 [31], X [37], 2 [43]
Post-Study System Usability Questionnaire (PSSUQ)	3	1 [6], 1 [7], 12 [31]
Attrakkdiff	2	341 [14], 2 [17]
Modular evaluation of key Components of User Experience (meCUE)	1	12 [14]
Usefulness, Satisfaction and Ease of use questionnaire (USE)	2	1 [6], 1 [17]
User Experience Questionnaire (UEQ)	2	200 [14], 1 [17]
Computer System Usability Questionnaire (CSUQ)	3	1 [4], 2 [6], 1 [7]
Software Usability Measurement Inventory (SUMI)	1	1 [46]
Usability Metric for User Experience (UMUX)	1	1 [17]
Unified Theory of Acceptance and Use of Technology (UTAUT)	1	2 [7]

in its original form, being the non-standardized questionnaire to be applied the most.

One study [14] focused particularly on three different standardized questionnaires: AttrakDiff, UEQ (User Experience Questionnaire) and meCUE (Modular Evaluation of key Components of User Experience), with the purpose of determining how they were applied in the past for Ambient Intelligence (AmI) and ubiquitous computing. Apart from the standardized questionnaires, one published non-standardized questionnaire was found applied by primary studies three times [31], seven were found applied twice in the primary studies [4, 17, 43, 46], and eighteen were found applied only once in the primary studies [4, 7, 17, 34, 46].

Plenty of secondary studies encountered methods and techniques that involved the researchers having direct contact with end user for evaluating the QinU. Some of them performed direct interviews to collect the information they wanted to report on, asking predefined questions regarding the software they wanted to evaluate [4, 6, 7, 9, 17, 20, 29, 31, 33, 34, 36, 37, 43, 46], while others took into consideration people's opinion through user feedback [9, 29], listening to the users' comments and thoughts after they interacted with the applications in a free manner, and some applied the think aloud protocol [7, 17, 29, 31], where the people involved use the software that is being evaluated while describing, out loud, their actions (and expectations regarding what will come from it).

A number of secondary studies found authors that used focus groups to evaluate the software quality [9,17,20,29,31,33,34,46], while some also indicated the observation method [4, 7, 17, 18, 20, 29, 34, 46].

Four secondary studies found the use of log data in primary studies [11, 17, 33,34,46], and in such cases, they were used for calculating metrics, even though they were evaluating the QinU in different systems. Two of these studies focused on the assessment of the QinU of smart environments for the elderly [33, 46], one [34] aimed to evaluate QinU characteristics in wearable devices, such as smartwatches, and another [17] had the goal of analyzing multi-touch systems in general. In all these cases, the log data came from the interaction of the end users themselves with the application that was being evaluated.

We noticed that only one study [19] pointed out the use of simulations applied in different platforms for different contexts of use. Most of them provided mathematical simulations, simulating "energy cost, hardware frequency rate, and computation time" of the scheme to be validated, but some other examples were of simulating "a resource description framework for heterogeneous IoT devices". There was no mention of any primary studies performing agent-based simulations for replacing human-computer interaction.

Some secondary studies pointed out evaluation methods and techniques based on different testing performances, which included real testbeds and prototypes [19], automated tests [29], and experimental protocols [32, 33, 39].

Finally, we defined a category "others" in Table 1 to represent some specific findings. First, to represent some methods and techniques for evaluation found only once in a secondary study: card sort [9], which is a method to help the researcher perceive how others categorize information (guaranteeing a data architecture that matches the users' expectations), and randomized controlled trials [36], which are techniques that balance the characteristics of the participants between the trial groups to allow differences in the intervention results. Similar to this last technique, four studies are identified in [32] as applying randomized clinical trials. Another study [19] found the use of formal techniques, employed to model complex systems as mathematical entities, and used what the authors called "design techniques" for evaluation, in which the primary studies aim to provide new approaches or frameworks. Task completion was also quoted [31] as another method for evaluation that considers whether or not the users were able to perform (with adequate standards) certain tasks within a defined period of time. Finally, several other approaches were quoted in the studies, not really precising which technique or method was applied [11, 19, 20]. For instance, [11] mentioned that the identified approaches in the primary studies worked on visual data anonymization methods to try "to retain all the informative richness of visual data acquired with RGB cameras" and on security based issues for user authentication and data encryption. Another one [20] mentioned the use of experimentation field tests in general to collect data.

It is noticeable that for most of the secondary studies there was a higher number for methods and techniques than for primary studies, which is caused by many of them combining different ways of evaluating the QinU of the application

under analysis. One example is when a secondary study [29] pointed out that two primary studies applied focus groups and interviews together.

We cannot quantify all the information from the primary studies found in the secondary ones, as some of them did not provide all the references and it is likely that there is an overlap of information. Besides, even though some secondary studies went into detail about their findings, some of them did not disclose specific numbers for the data taken from primary studies (for example, how many of them applied the same method or technique for assessing software quality or how many measured a specific QinU characteristic).

4.2 RQ2: What are the most evaluated types of systems?

The most evaluated type of systems were smart environments/AmI [6, 7, 9, 11, 14, 15, 18, 33, 36, 43], with ten studies (47,61%), as seen in Fig. 2, followed by IoT [19, 20, 32, 34, 37, 39, 46] with seven studies (33,33%). The other software categories were evaluated by only one study each, those being cyber-physical systems [4], multi-touch systems [17], eGovernment [29], and eHealth [31].

Most of the secondary studies about smart environments/AmI focused on Ambient Assisted Living (AAL) [6,7,9,11,14,15,33,36]. However, one researched specifically ubiquitous healthcare [43] and another one smart learning environments [18]. Among the seven studies that included QinU evaluations for IoT, two of them dealt with navigation apps: one of them [20] involved mobile safety alarms with GPS, RFID tags and readers, product design assessment for safe navigation and more. The other study [37] focused on indoor navigation apps for people with mobility disabilities. Three other studies about IoT investigated wearable devices (such as smartwatches, wristbands or neckwear) for different reasons: tracing people and objects regarding hospital-like scenarios involving medical teams and patients [19], improving medication adherence [32], or analyzing physical activities [34]. One study focused on assistive technology for older adults [46] and another on IoT in general [39].

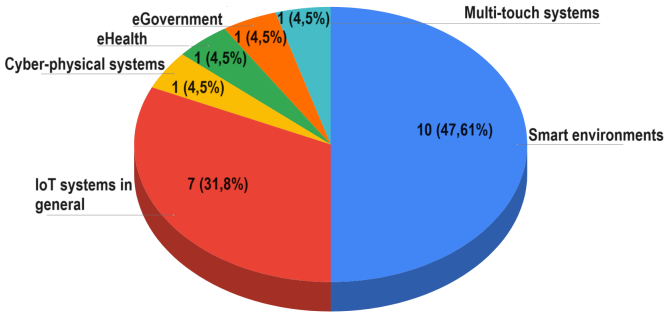


Fig. 2: Types of systems evaluated in the secondary studies.

Table 3: QinU characteristics found in the primary studies (#).

Ref.	#	Usability	Effectiveness	Efficiency	Satisfaction	Freedom from risk	Context coverage
[4]	24	12	5	3	3	1	-
[6]	44	44	-	-	-	-	-
[7]	35	35	-	-	-	-	-
[9]	21	21	-	-	-	-	-
[11]	63	40	-	-	6	17	-
[14]	553	553	-	-	-	-	-
[15]	21	15	17	17	17	14	13
[17]	65	65	13	11	2	-	-
[18]	15	7	-	-	3	-	-
[19]	146	-	-	X	-	X	-
[20]	29	29	-	-	-	-	-
[29]	22	22	-	-	-	-	-
[31]	133	133	-	-	-	-	-
[32]	9	-	6	2	1	3	-
[33]	34	-	34	-	-	-	-
[34]	111	16	107	-	20	-	-
[36]	8	-	8	-	-	-	-
[37]	51	51	-	-	-	-	-
[39]	12	4	2	2	2	5	-
[43]	10	10	-	-	-	-	-
[46]	31	-	31	-	-	-	-

4.3 RQ3: Which quality characteristics were the most evaluated?

All QinU characteristics as well as usability were found to be evaluated in the secondary studies, as seen in Table 3. It is usual for a study to evaluate more than one QinU characteristic at once. Usability was evaluated the most, being found in sixteen secondary studies (76,19%) [4,6,7,9,11,14,15,17,18,20,29,31,34,37,39,43]. This result is expected, considering that usability is the focus of questionnaires, which were the most commonly used method for evaluation (see Section 4.1). Effectiveness was the second most mentioned, with nine studies [4, 15, 17, 32–34, 36, 39, 46], followed by satisfaction, with eight studies [4, 11, 15, 17, 18, 32, 34, 39]. Then, tied with six studies each, efficiency [4, 15, 17, 19, 32, 39] and freedom from risk [4, 11, 15, 19, 32, 39]. Lastly, context coverage was found to be assessed by only one of the secondary studies [15].

It is noticeable that many studies apply questionnaires for evaluating usability (and, as a consequence, effectiveness, efficiency and satisfaction as defined

by ISO 9241-11 [21]) since the Human-Computer Interaction domain has been investigating usability issues for a long time.

Most of the studies aimed at evaluating the main QinU characteristics, but some had the goal of evaluating ones that were considered as subcharacteristics, according to the ISO/IEC 25010 standard [26], such as usefulness, acceptability [11, 18, 46], which fall under the satisfaction umbrella, and reliability [39], safety [11], security [11], which are considered a subcharacteristic of freedom from risk. It is important to follow a pattern to guarantee that studies have the goal of evaluating the same characteristics, as "it is evident that the concepts of UX and usability are not the same among the authors" [17]. However, few of the secondary studies extracted information about alignment with quality standards. One of them [15] pointed out that five primary studies followed ISO/IEC 25010 [26], four followed ISO/IEC 9126 [23], two followed ISO/IEC 25012 [24], and lastly, one followed ISO/IEC 14598 [22]. Another [39] classified the selected primary studies according to the characteristics and subcharacteristics they evaluated, splitting them between system/software product quality and QinU, according to ISO/IEC 25000 [25]. Moreover, none of the analyzed studies followed a standard such as ISO/IEC 25022 [27], which brings measures for the said characteristics to be assessed. One of the secondary studies [29] suggested that future researches should take ISO/IEC 25010:2011 [26] into consideration, as it defines usability in a context of use.

5 Threats to Validity

The analyses of threats to the validity of this study have been based on Petersen [41], as follows:

- *Descriptive validity*, related to the gathering of data, which includes recording, storing, and analyzing the information that is being reviewed. To mitigate this threat, a Google Form was created for extracting data according to the established RQs, and that information was then accessed in a spreadsheet on Google Sheets. Moreover, all quantitative analysis was double checked by one or two peers;
- *Theoretical validity*, which involves the identification of the studies to be analyzed and also the data extraction. It is possible that some studies have not been included in this review because the search was performed on two databases, which might provide incomplete results. However, Scopus and Web of Science are recognized for indexing different conferences and journals, and both have been widely used in reviews [35]. To mitigate the risk of bias interpretation, all data extraction was peer-reviewed. However, since this step involves human judgment, the threat cannot be completely eliminated;
- *Generalizability validity* worries about how generalizable the obtained results can be in a wider scenario (within an institution or between several different groups or organizations). As a tertiary study, this factor depends on the generalizability of the 21 secondary studies performed. Another issue

that might impact our work is that since not all of the secondary studies mention their references, it is very likely that there are overlaps of information regarding the analyzed QinU evaluations. Therefore, we prefer to be cautious and say that we cannot guarantee the generalizability of the results in different organizations, but since we used the databases available for our institution, it can be considered generalizable in our institution;
- *Repeatability validity* requires that all methods are established and reproducible. We consider this threat under control, since the protocol and all the analysis procedures that took place were well documented, with the detailed process [2] and replication package [1] being available on HAL[6].

6 Final remarks

This paper presents a tertiary review about evaluating the Quality in Use of smart environments, pervasive systems and intelligent applications. Our goal was to identify how this evaluation has been done considering, in particular, the case of smart environments inhabited by elderly or disabled people.

By carrying out this study, contrary to our expectation, we found that most of the assessment approaches to evaluate the QinU of smart environment applications use classic methods and techniques requiring that users interact directly with the application and answer usability questionnaires/interviews. That means the user should have a live experience in these environments before answering questionnaires or attending interviews to offer their feedback. We believe that carrying out these experiences with elderly or people with disabilities is not adequate. There is, therefore, an urgent need to develop tools that can carry out these evaluations in an automated way. Our ongoing work looks to address this gap with the development of an approach to evaluate the QinU of smart environment applications without having to involve end users, and instead involve artificial intelligence and agent-based modeling and simulation as a step that would come before any end user evaluation [3].

Acknowledgments

The present research work is partially supported by the Hauts-de-France Region and the REUNICE (EUNICE Research) project funded by the European Union and the CNRS. C. Tirnauca's work is partially supported by the project PID2022-139237NB-I00 financed by MICIU/AEI/10.13039/501100011033 and FEDER, UE.

References

1. Angeloni, M.P.C., Duque Medina, R., Marçal de Oliveira, K., Grislin-Le Strugeon, E., Tirnauca, C.: Initial results for a tertiary study on quality in use evaluation of smart environment applications (Mar 2024), https://hal.science/hal-04506543

[6] https://hal.science/

2. Angeloni, M.P.C., Duque Medina, R., Marçal de Oliveira, K., Grislin-Le Strugeon, E., Tirnauca, C.: Research protocol and results for a tertiary study on quality in use evaluation of smart environment applications (Mar 2024), https://hal.science/hal-04507278
3. Angeloni, M.P.C., Marçal de Oliveira, K., Grislin-Le Strugeon, E., Duque, R., Tirnauca, C.: A review on quality in use evaluation of smart environment applications: What's next? In: Comp. Proc. of the 2023 ACM SIGCHI Symposium on Engineering Interactive Computing Systems. p. 9–15 (2023). https://doi.org/10.1145/3596454.3597177
4. Apraiz, A., Lasa, G., Mazmela, M.: Evaluation of user experience in human–robot interaction: A systematic literature review. Int. J. of Social Robotics **15**(2), 187 – 210 (2023)
5. Assila, A., Marçal de Oliveira, K., Ezzedine, H.: Standardized Usability Questionnaires: Features and Quality Focus. Elec. J. of Comp. Sc. and Inf. Tech. **6**(1), 15–31 (Dec 2016)
6. Bastardo, R., Martins, A.I., Pavão, J., Silva, A.G., Rocha, N.P.: Methodological quality of user-centered usability evaluation of ambient assisted living solutions: A systematic literature review. Int. J. of Environmental Research and Public Health **18**(21) (2021)
7. Bastardo, R., Pavão, J., Rocha, N.P.: A scoping review of the inquiry instruments being used to evaluate the usability of ambient assisted living solutions. In: Proc. of the 15th INSTICC Int. Joint Conf. on Biomedical Eng. Systems and Tech. - vol. 5: HEALTHINF. pp. 320–327 (2021)
8. Carbo, J., Sanchez-Pi, N., Molina, J.M.: Agent-based simulation with netlogo to evaluate ambient intelligence scenarios. J. of Simulation **12**(1), 42–52 (2018). https://doi.org/10.1057/jos.2016.10
9. Choukou, M.A., Shortly, T., Leclerc, N., Freier, D., Lessard, G., Demers, L., Auger, C.: Evaluating the acceptance of ambient assisted living technology (aalt) in rehabilitation: A scoping review. Int. J. of Medical Informatics **150** (2021)
10. Cioroaica, E., Buhnova, B., Kuhn, T.: Predictive simulation within the process of building trust. In: 2022 IEEE 19th Int. Conf. on Software Architecture Companion, ICSA-C 2022. pp. 47–48 (2022). https://doi.org/10.1109/ICSA-C54293.2022.00017
11. Colantonio, S., Jovanovic, M., Zdravevski, E., Lamesky, P., Tellioglu, H., Kampel, M., Florez-Revuelta, F.: Are active and assisted living applications addressing the main acceptance concerns of their beneficiaries? preliminary insights from a scoping review. In: Proc. of the 15th ACM Int. Conf. on Pervasive Tech. Related to Assistive Environments. pp. 414–421 (2022)
12. Costal, D., Farré, C., Franch, X., Quer, C.: How tertiary studies perform quality assessment of secondary studies in software engineering. XXIV Iberoamerican Conf. on Software Eng. (ESELAW@CIbSE) (2021). https://doi.org/https://doi.org/10.48550/arXiv.2110.03820
13. Demiris, G., Hensel, B.K., Skubic, M., Rantz, M.: Senior residents' perceived need of and preferences for "smart home" sensor technologies. Int. J. of Technology Assessment in Health Care **24**(1), 120–124 (2008). https://doi.org/10.1017/S0266462307080154
14. Díaz-Oreiro, I., López, G., Quesada, L., Guerrero, L.A.: Ux evaluation with standardized questionnaires in ubiquitous computing and ambient intelligence: A systematic literature review. Advances in Human-Computer Interaction **2021** (2021)
15. Erazo, L., Erraez, J., Cedillo, P., Illescas Illescas-Peña, L.: Quality Assessment Approaches for Ambient Assisted Living Systems: A Systematic Review, pp. 421–439. Springer International Publishing, Cham (2020)

16. Felber, N.A., Tian, Y.J.A., Pageau, F., Elger, B.S., Wangmo, T.: Mapping ethical issues in the use of smart home health technologies to care for older persons: a systematic review. BMC Medical Ethics **24**(1) (2023). https://doi.org/10.1186/s12910-023-00898-w
17. Filho, G.E.K., Guerino, G.C., Valentim, N.M.C.: A systematic mapping study on usability and user experience evaluation of multi-touch systems. In: Proc. ACM 21st Brazilian Symp. on Human Factors in Computing Systems (2022)
18. Gambo, Y., Shakir, M.Z.: Review on self-regulated learning in smart learning environment. Smart Learning Environments **8**(1) (2021)
19. Haghi Kashani, M., Madanipour, M., Nikravan, M., Asghari, P., Mahdipour, E.: A systematic review of IoT in healthcare: Applications, techniques, and trends. J. of Network and Computer Applications **192** (2021)
20. Holthe, T., Halvorsrud, L., Karterud, D., Hoel, K.A., Lund, A.: Usability and acceptability of technology for community-dwelling older adults with mild cognitive impairment and dementia: A systematic literature review. Clinical Interventions in Aging **13**, 863 – 886 (2018). https://doi.org/10.2147/CIA.S154717
21. ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDT) s- Part 11 Guidance on usability (1998)
22. ISO/IEC: 14598-5:1998. information technology — software product evaluation — part 5: Process for evaluators (1998), https://www.iso.org/standard/24906.html
23. ISO/IEC: 9126. software engineering — product quality (2001), https://www.iso.org/standard/22749.html
24. ISO/IEC: 25012:2008. software engineering — software product quality requirements and evaluation (square) — data quality model (2008), https://www.iso.org/standard/35736.html
25. ISO/IEC: 25000: Systems and software quality requirements and evaluation (square) — guide to square (2011), https://www.iso.org/standard/64764.html
26. ISO/IEC: 25010:2011. systems and software engineering — systems and software quality requirements and evaluation (square) — system and software quality models (2011), https://www.iso.org/standard/35733.html
27. ISO/IEC: 25022:2016. systems and software engineering — systems and software quality requirements and evaluation (square) — measurement of quality in use (2016), https://www.iso.org/fr/standard/35746.html
28. Kitchenham, B., Charters, S.: Guidelines for performing structural literature reviews in software engineering (version 2.3). Tech. Report, Keele Univ. and Univ. of Durham (2007)
29. Lyzara, R., Purwandari, B., Zulfikar, M.F., Santoso, H.B., Solichah, I.: E-government usability evaluation: Insights from a systematic literature review. In: Proc. ACM 2nd Int. Conf. on Software Engineering and Information Management. p. 249–253 (2019)
30. Mao, C., Chang, D.: Review of cross-device interaction for facilitating digital transformation in smart home context: A user-centric perspective. Adv. Engineering Informatics **57**, 102087 (2023). https://doi.org/https://doi.org/10.1016/j.aei.2023.102087
31. Maramba, I., Chatterjee, A., Newman, C.: Methods of usability testing in the development of ehealth applications: A scoping review. Int. J. of Medical Informatics **126**, 95 – 104 (2019). https://doi.org/10.1016/j.ijmedinf.2019.03.018
32. Marengo, L.L., Barberato-Filho, S.: Involvement of human volunteers in the development and evaluation of wearable devices designed to improve medication adherence: A scoping review. Sensors **23**(7) (2023)

33. Maresova, P., Krejcar, O., Barakovic, S., Barakovic Husic, J., Lameski, P., Zdravevski, E., Chorbev, I., Trajkovik, V.: Health-related ict solutions of smart environments for elderly-systematic review. IEEE Access **8**, 54574 – 54600 (2020)
34. McCallum, C., Rooksby, J., Gray, C.: Evaluating the impact of physical activity apps and wearables: Interdisciplinary review. JMIR mHealth and uHealth **6**, e58 (03 2018). https://doi.org/10.2196/mhealth.9054
35. Motta, R.C., de Oliveira, K.M., Travassos, G.H.: A conceptual perspective on interoperability in context-aware software systems. Information and Software Technology **114**, 231–257 (2019). https://doi.org/https://doi.org/10.1016/j.infsof.2019.07.001
36. Moyle, W., Murfield, J., Lion, K.: The effectiveness of smart home technologies to support the health outcomes of community-dwelling older adults living with dementia: A scoping review. Int. J. of Medical Informatics **153** (2021). https://doi.org/10.1016/j.ijmedinf.2021.104513
37. Nasr, V., Zahabi, M.: Usability evaluation methods of indoor navigation apps for people with disabilities: A scoping review. In: Proc. of the 2022 IEEE Int. Conf. on Human-Machine Systems. pp. 1–6 (2022)
38. Ntoa, S., Margetis, G., Antona, M., Stephanidis, C.: Uxami observer: An automated user experience evaluation tool for ambient intelligence environments. Advances in Intelligent Systems and Computing **868**, 1350–1370 (2018). https://doi.org/10.1007/978-3-030-01054-6_94
39. Paiva, J.O., Andrade, R.M., Carvalho, R.M.: Evaluation of non-functional requirements for iot applications. In: Int. Conf. on Enterprise Information Systems, ICEIS - Proceedings. vol. 2, p. 111 – 119 (2021)
40. Pavlovic, M., Kotsopoulos, S., Lim, Y., Penman, S., Colombo, S., Casalegno, F.: Determining a framework for the generation and evaluation of ambient intelligent agent system designs. Adv. in Intelligent Systems and Computing **1069**, 318–333 (2020). https://doi.org/10.1007/978-3-030-32520-6_26
41. Petersen, K., Gencel, C.: Worldviews, research methods, and their relationship to validity in empirical software engineering research. In: 23rd Int. W. on Software Measurement and the 8th Int. Conf. on Software Process and Product Measurement. pp. 81–89 (10 2013). https://doi.org/10.1109/IWSM-Mensura.2013.22
42. Petticrew, M., Roberts, H.: Systematic reviews in the social sciences: A practical guide. Blackwell Publishing Ltd, Oxford, UK (2006). https://doi.org/10.1016/B978-0-12-374708-2.00014-0
43. Saleemi, M., Anjum, M., Rehman, M.: Ubiquitous healthcare: a systematic mapping study. J. of Ambient Intelligence and Humanized Computing **14**(5), 5021 – 5046 (2023)
44. Singh, V.K., Singh, P., Karmakar, M., Leta, J., Mayr, P.: The journal coverage of web of science, scopus and dimensions: A comparative analysis. Scientometrics **126**(6), 5113–5142 (jun 2021)
45. Tricco, A., Langlois, E., Straus, S., for Health Policy, A., Research, S., Organization, W.H.: Rapid reviews to strengthen health policy and systems: a practical guide. World Health Organization (2017)
46. Tónay, G., Pilissy, T., Tóth, A., Fazekas, G.: Methods to assess the effectiveness and acceptance of information and communication technology-based assistive technology for older adults: A scoping review. Int. J. of Rehabilitation Research **46**(2), 113–125 (2023). https://doi.org/10.1097/MRR.0000000000000571