Journal on Interactive Systems, 2023, 14:1, doi: 10.5753/jis.2023.3279

Usability and User Experience of Multi-touch Systems: A Systematic Mapping Study and Benchmark

Guilherme Konopatzki Filho [**Universidade Federal do Paraná**| ***guilherme.ekf@gmail.com***]
Guilherme Corredato Guerino [**Universidade Estadual de Maringá** | ***guilherme.guerinosi@gmail.com***]
Natasha M. C. Valentim [**Universidade Federal do Paraná** | ***natasha@inf.ufpr.br***]

Abstract Each software context has its specificities. One specfic type of context is the multi-touch, where two or more touches are recognized by the software at the same time. Still in this context, User eXperience (UX) and Usability are relevant criteria related to multi-touch systems quality, and evaluation technologies are used to assess this quality. But which are these technologies? This is the central question that the Systematic Mapping Study (SMS) seeks to answer. Besides this question, another 18 sub-questions were addressed to find out more about the peculiarities of the technologies and how they can affect the state of the art of evaluating multi-touch systems. This SMS returned 622 papers, which were analyzed using two filters. Finally, 65 papers had their data extracted through the 18 sub-questions. These extractions raised information such as the software used, data collection methods, and aspects evaluated by the technologies. Through these data, we noticed an absence of technologies explicitly built for the multi-touch context. Other gaps were also perceived, such as the need for technologies that jointly assess quantitative and qualitative data; and technologies that focus on jointly evaluating Usability and UX. Perceiving the lack of synthesized content and characteristics about the questionnaires, we also performed a benchmark over the questionnaires identified in the SMS to serve as a future guide when applicators must choose the better technology for their context.

Keywords: *Multi-touch, Usability, User Experience, Evaluation Technologies*

1 Introduction

The interaction with electronic devices has been changing since its emergence. An intuitive way to interact with these devices (e.g., smartphones, tablets, tabletop) is the direct manipulation (Shneiderman, 1981). In 1983, Shneiderman (Shneiderman, 1981) first introduced the term direct manipulation to describe the idea of interaction with the Graphical User Interface (GUI) using a pointer through a mouse. Today, with the invention of the touch screen, we can just put our hands on the screen to interact with it (Damaraju et al., 2013). On these screens, a significant part of the interaction can be classified as multi-touch, where more than one touch is recognized simultaneously in the system. Like any device, corrections and improvements in multi-touch systems must be made over time to evolve and make better use of them.

In this way, Usability and User eXperience (UX) evaluations help control and verify the quality of the software being developed and used. Usability is described as ”the degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” according to the ISO/IEC 25010 (2011). Therefore, Usability evaluation helps to improve software quality by assessing aspects that determine how the user will perceive the interaction. On the other hand, UX is defined by ISO 9241-210 (2019) as the ”perceptions and responses of the person that result from the use and/or early use of a product, system or service”. In this way, UX is about the users’ feelings, emotions, and judgments when using a system (Hassenzahl and Tractinsky, 2006). Also is said by Hassenzahl and Tractinsky (2006) that: “UX in the sense of a positive HCI would, thus, focus on how to create outstanding quality experiences rather than merely preventing usability problems”. This definition infers that UX is something beyond Usability because it aims to provide ”quality experiences”, while Usability aims to solve behavioral problems of the system. Therefore, the concept of UX evaluation used in our paper is the evaluation of the experiences, emotions and feelings linked directly to the users.

Usability and UX evaluation perform an important role in understanding how to improve user interaction with any system, including multi-touch systems. Therefore, our goal is to identify the technologies used to evaluate Usability and UX in multi-touch systems. The research question that guided this process was ”What technologies are used to evaluate the Usability and UX of software that uses multi-touch-based interaction?”. It is important to emphasize that the concept of technology used in this study is the one used by Santos et al. (2012), which defined ”technology” as a generalization for metrics, tools, methodologies, techniques, among others. For this identification, we conducted a Systematic Mapping Study (SMS), which aims to characterize the state of the art of a given topic. To this end, our SMS also extracted data about multi-touch software function, quality criteria, aspects of Usability and UX, number of fingers used to interact with the multi-touch application, and data collection methods, among others.

This SMS was based on Kitchenham and Charters (2007) protocol and comprehend search string, goals, data extraction strategy, research questions, inclusion and exclusion criteria, and data analysis. We filtered returned publications and extracted data from 65 papers following this protocol, identifying 29 different technologies. The discoveries after analysis show that most of the technologies do not evaluate Usability and UX together. The analysis also reveals that most of the technologies are generic, not taking into account the specificities of the multi-touch context. Finally, our results demonstrate that most of these technologies are created for the study and not based on existing ones, showing a lack of standardization among the technologies used.

We perceived a significant difficulty finding synthesized information about the content and characteristics of questionnaires that evaluate usability and UX in the multi-touch context. This difficulty may also exist for other researchers and interested parties. Therefore, we performed a benchmark of the main usability and UX evaluation questionnaires. These questionnaires are generic, that is, are used to evaluate any kind of system. However, in the papers found in our SMS, we identified the use of these questionnaires in the multi-touch context. Thus, we analyzed characteristics of these questionnaires, like the number of questions, criteria, scales, if use questions for each aspect, data collection method, aspects evaluated and if they have an automated version. We hope that this synthesis and analysis can facilitate the process of choosing the most appropriate assessment technology for the context of future research.

This paper is organized as follows: Section 2 presents this research background; Section 3 shows the structure of our SMS; Section 4 presents the SMS results; Section 5 shows the benchmark realised; Section 6 shows the discussion of results; Section 7 presents the threats to validity; Section 8 presents conclusions and future work.

2 Background

2.1 Natural user interfaces

Natural User Interfaces (NUI) was popularized in 2006 due to a conference where Jefferson Hen presented the results of his research on multi-touch interfaces. However, research covering this topic has been conducted since the 1990s (Glonek and Pietruszka, 2012). NUIs feature interactive technologies that help communicate with operating systems (Wigdor and Wixon, 2011). The natural property does not refer to the interface but to how users interact and their feelings throughout the experience. According to Wigdor and Wixon (2011), a natural user interface with good design should be developed to be perceived as an extension of the user's senses.

Among the ways of interaction that this type of interface can benefit from, Vetere et al. (2014) and Fernandez et al. (2016) highlight gestures, look, voice and multi-touch. Gestures are body movements that can be detected by systems for the desired interaction. Its popularization was marked by the launch of devices such as Kinect[1], being mainly used for 3D games and manipulation. Eye-based systems are represented by Eye Trackers technologies, where eye tracking occurs to perform interactions. Technologies like these have been used in cases of people with varying levels of paralysis, like Stephen Hawking. Voice-based interaction has become popular through voice assistants and devices such as Alexa[2], which can do various tasks, from creating a shopping list, to controlling a smart home.

Multi-touch is one of the most common forms of NUIs, being present in most smartphones today. There are several examples of routine uses, such as the pinch gesture when applying zoom in map applications or the use of 2 or more fingers in mobile games. This study area also arouses the interest of HCI and Software Engineering (SE) researchers. Several studies explore the implementation of multi-touch interfaces when targeted at different specific groups, such as children (Zaharias et al., 2013), drivers (Pfeiffer et al., 2010), students (Fu et al., 2010), editors (Damaraju et al., 2013), among others. Besides, it is interesting to notice that the industry of smartphones, the primary representative devices of multitouch interaction, has been growing in the last few years. It is estimated that this movement should reach US$ 493.13 billion in 2026, according to data from Market Data Forecast (2022).

2.2 Related Work

Some secondary studies approach Usability and UX evaluation about touch interface. For example, Fleh et al. (2018) performed an SMS on touch classification of social touch gesture recognition. From the 938 papers returned, after three filters, 49 were extracted. This study had four main questions regarding methods and algorithms identification, main factors that affect the study results, sensors and interfaces used, and improvement of research results. In their search string, terms such as accuracy, performance, and efficiency were used, which are known to be Usability aspects. Many facets were used to create a filter distribution of the papers, including classification methods, research type, touch surface, touch recognition, and sensor type. The results obtained through the touch surface show that the most used touch type is the mannequin arm, while the touch recognition of facet explains the gesture recognition that most studies use.

In Dodd et al. (2017), a Systematic Literature Review (SLR) was made to understand state of the art about the development of interfaces specified for elderly users. The UX concern about this specific group is expressed through the study's central questions, which treat challenges experienced by elderly users while using an interface and the respective solutions for these challenges. Aspects such as intuitiveness and attention are addressed. Intuitiveness is one of the aspects considered UX by some researchers (Schürmann et al., 2015) and Usability by others (Liang et al., 2011). This study shows that every paper analyzed brings a proposed solution for the challenges addressed. Thus, the authors map and group the solutions for each challenge encountered.

Some secondary studies were also identified when we searched for a specific context of multi-touch systems. In Nayebi et al. (2012), the authors have examined several studies to identify the main Usability challenges in the mobile context. In their research was observed in the literature that the questionnaires and methods made for mobile Usability assessment and measurement do not take into consideration the

[1] https://docs.microsoft.com/pt-br/windows/apps/design/devices/kinect-for-windows

[2] https://developer.amazon.com/pt-BR/alexa

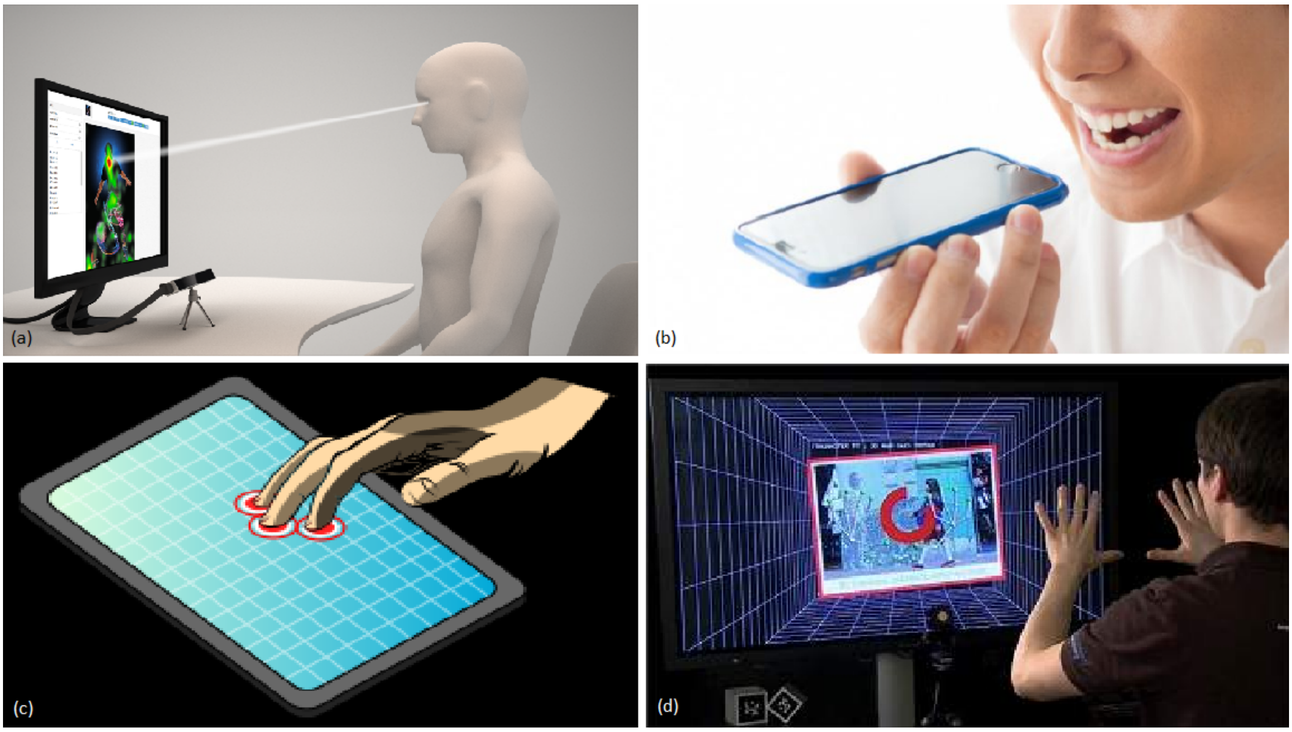

Figure 1. Interaction types used in NUIs: (a) Gaze. (b) Voice. (c) Multi-touch. (d) Gestures.

user interface features provided by the current mobile operating systems that are gaining popularity, being one of them the multi-touch gestures, such as tap, flick, and pinch. Part of the results shows that there is no scientific research focusing on the newest requirements of the mobile user interfaces. Also, a need to develop a questionnaire as a field study methodology is evidenced in considering new mobile operating system needs.

Guerino and Valentim (2020) performed an SMS to identify Usability and UX evaluation technologies used by researchers and developers in software with Natural User Interfaces. The study's central question sought to identify what technologies are used to evaluate the Usability and UX of software with NUI. The search was performed in Scopus, IEEExplore, ACM Digital Library, and Engineering Village. Their findings showed that 14 of the 119 found technologies evaluated the Usability and/or UX of software that used multi-touch-based interfaces. This represents 11.76% of the found technologies. Besides, none of these 14 technologies are specific for the multi-touch context.

Some works evaluated Usability and UX, while others challenge multi-touch interfaces, and others both. However, none of them deals with the specifics of Usability and UX in the multi-touch context, a gap that our work intends to fill with the SMS shown in the sequential section.

3 SMS Structure

Seeking to build a path based on evidence, the methodology that most represents the research proposal is presented by Mafra and Travassos (2006). This methodology allows the use and synthesis of knowledge from primary and secondary studies. Primary studies serve to directly test hypotheses, while secondary studies help in a better understanding of a particular area of research. The study by Mafra and Travassos (2006) is a study that expands the methodology proposed by Shull et al. (2001), which exposes feasibility, observation, and case studies to evaluate technology from its proposal to its transfer to industry. This expansion occurs by demonstrating the advantages of conducting a secondary study before the primary studies. These advantages can be summarized by knowing the area in focus better and having more information about the proposition itself.

SMS is a kind of secondary study where the goal is to define the state of the art of a determined research topic (Kitchenham and Charters, 2007). Our choice to perform an SMS was based on the reliability and organization of the information collected and synthesized by this method, which can be used in the future to carry out other SMS and SLR. Our SMS followed the guidelines proposed by Kitchenham and Charters (2007), where the process is divided into three phases: planning, execution, and reporting. The planning phase defines a mapping protocol, research questions, reliable sources to find the data, a search string, and inclusion/exclusion criteria. The execution phase encompasses performing searches on data sources chosen previously and filtering the papers found through two filters. From this selection, the filtered papers have their data extracted and analyzed. Finally, in reporting, results are organized and shared.

3.1 Phase 1: Planning

3.1.1 Goal.

The SMS goal was based on Goal-Question-Metric (GQM) (Basili and Rombach, 1988) paradigm and is described in Table 1.

Table 1. SMS goal.

Analyze	scientific publications
For the purpose of	to categorize
With respect to	technologies that evaluate the Usability and/or UX of software that use multi-touch-based interaction
From the point of view of	Human-Computer Interaction (HCI) researchers
In context	publications available from ACM, Engineering Village and SCOPUS

3.1.2 Research Questions.

The main question of our SMS is "What technologies are used to evaluate the Usability and UX of software that uses multi-touch-based interaction?". Of this question we unfolded 3 other main questions, that are Q1, Q2 and Q3. Santos et al. (2012) defined "technology" as a generalization for metrics, tools, methodologies, and techniques. Others 18 sub-questions (SQs) were defined. These SQs address topics about the multi-touch systems evaluated in the papers, specific technical features of the evaluation technologies, and empirical evaluations. These SQs are present in Tables 2 and 3.

3.1.3 Data Sources.

We choose the data sources *ACM Digital Library*[3], *Engineering Village*[4] and *Scopus*[5] because they: (i) return a relevant number of papers; (ii) allow the exact search string to be used; (iii) provide an efficient search engine, and (iv) are sources known to contain relevant papers in the area of HCI.

3.1.4 Search String.

The PICOC criterion (Population, Intervention, Comparison, Outcome, and Context) (Kitchenham and Charters, 2007) was applied to obtain the keywords that would be used in the search string. This search string was used to perform the search through the full papers. In our SMS, keywords for Context and Comparison were not used in the string since the goal of SMS was not to compare evaluation technologies but to characterize them. Therefore, PICOC was defined as follows:

- (P)opulation: multi-touch based interaction systems;
- (I)ntervention: technologies that evaluate Usability and/or UX used in the software development process that uses multi-touch-based interaction;
- (C)omparison: not applicable;
- (O)utcome: evaluation of the Usability and/or UX of the multi-touch interaction system;
- (C)ontext: not applicable;

Table 4 shows the English terms that compose the search string, which was divided from the definition of population, intervention, and outcome of the PICOC criterion.

[3]https://dl.acm.org/
[4]https://www.engineeringvillage.com/home.url
[5]https://www.scopus.com/search/form.uri?display=basic

3.1.5 Inclusion Criteria

- IC1. Publications that present technologies that evaluate Usability and/or UX in the software development process that uses multi-touch interaction;
- IC2. Publications that describe experimental studies of technologies used to evaluate Usability and/or UX in the development process that uses multi-touch based interaction;

3.1.6 Exclusion Criteria

- EC1. Publications that do not meet the above criteria;
- EC2. Publications that do not have content available for reading and data analysis were not selected (especially in cases where studies are paid for or not made available by search engines);
- EC3. Publications that have a language other than English and Portuguese were not selected;
- EC4. Publications that are part of the gray literature, such as technical reports and work in progress, were not selected;
- EC5. Publications that have already been included in another search engine (duplicates) were not selected.

3.2 Execution

3.2.1 Publication Selection.

We searched data sources to identify the publications in July 2021. Three researchers carried out the selection and extraction of data shown in our SMS. Kitchenham and Charters (2007) declare that the use of two or more researchers to reduce bias and maintain research consistency is essential. In the first filter, the first researcher evaluated all papers returned based on title and abstract and set a classification of excluded or included based on exclusion and inclusion criteria. Then, the second and third researchers performed the first filter separately. If differences were found in exclusion or inclusion decisions, researchers discussed resolving them. The paper was accepted for a more detailed reading if researchers did not reach a conclusion based on the title and abstract. For a paper rejection, a justification was always necessary.

In the 2nd and last filter, the first researcher read and classified all papers approved in the first filter and extracted data. Then, the second and third researchers verified the excluded papers and justifications and the included papers and their extractions. Then, from the same inclusion and exclusion criteria, researchers selected the papers to extract the data. Likewise, if the paper was excluded, a plausible justification was presented. To perform a collaborative SMS, the tool Porifera was used (Campos et al., 2022). The Fleiss Kappa for the first review among the researchers was of 0.6683, and in the second review the values were 0.6993. In both filters, the values are considered good, according to Altman (1990).

Table 5 shows that 622 papers were returned after applying the search string in the mentioned data sources. From the first filter, we selected 155 papers. In the second filter, we selected and extracted 65 papers, which can be seen in Tables 6 and 7.

Table 2. Research sub-questions from SQ1 to SQ12.

Sub-question (SQ)	Possible Answers
Evaluation Technology	
SQ1. What is the quality criterion of evaluation technology used?	Technology may have the following criteria: 1. Usability: technology aims to evaluate system Usability; 2. User eXperience: technology aims to evaluate system UX; 3. Both: technology aims to evaluate both Usability and UX.
SQ2. What aspects of Usability and/or UX does the technology evaluate?	Answers of SQ2 are subjective and each author defines what is the aspect cited. Evaluated aspects of Usability and/or UX can be user satisfaction, effectiveness, immersion, fatigue, pleasure, among others.
SQ3. Is the technology specific to multi-touch systems or to systems in general?	Technology can be: 1. Specific: Usability and/or UX evaluation technology is specific to multi-touch systems; 2. Generic: Usability and/or UX evaluation technology is not restricted to a specific type of software.
SQ4. Was the technology created for the study or is it based on an existing one?	Technology can be: 1. Existing: technology is based on an existing one; 2. Created: technology was created for the study reported in the paper.
SQ5. How does evaluation technology collect data from participants?	Answers to SQ2 are subjective and each paper have their own way to answer this. Its goal is to verify how users' responses were captured using the evaluation technology, such as Likert scale and open-ended questions.
SQ6. What are the characteristics of evaluation technology?	Answers to SQ6 are subjective and differ from paper to paper. SQ6 goal is to extract attributes of each technology, such as identifying questions used in case of questionnaires of interviews; identify metrics used in the case of technology-based on error rate; identify calculations performed in the case of a Usability test, among others.
SQ7. Does technology extract quantitative or qualitative data?	Technology can extract data: 1. Quantitative: if the analysis of evaluation was done quantitatively; 2. Qualitative: if the analysis of evaluation was done qualitatively; 3. Mixed: if the analysis of evaluation was done qualitatively and quantitatively.
Evaluated Multi-Touch System	
SQ8. What is the function of the multi-touch system?	This answer is subjective and differ from paper to paper. This question could have answers like navigation, chord reproduction, 3D object manipulation, switch settings, among others.
SQ9. Which application category does the multi-touch system?	Answers obtained in SQ9 are subjective and were identified during the papers reading. Examples of categories are health, education, urban mobility, daily facilities, among others.
SQ10. Was the software created for a specific group of people? Which groups?	Possible answers are: 1. Yes: the system was created to assist a specific group. Examples: blind, elderly, children, among others; 2. No: the system was not created for a specific group.
SQ11. How many fingers are used to interact with the multi-touch system?	This answer could include from 2 to 10 fingers, or could be unidentified, when the exact number of fingers could not be identified in the study.
SQ12. What is the gesture used in the multi-touch interaction?	This answer varies from paper to paper, and the gestures could be pinch, touch, rotation, among others.

Table 3. Research sub-questions from SQ13 to SQ18.

Evaluation Study of Assessment Technology	
SQ13. The evaluation technology was submitted through an evaluation study?	This answer could be answered simply with yes or no.
SQ14. In case of affirmative answer of SQ13, what was the type of the study realized?	The answers for this question differ for each paper, and possible answer include feasibility studies, case studies, observation studies, among others.
SQ15. How many participants were there?	This question gathered the number of participants present in each evaluation study.
Evaluation Study of Multi-Touch System	
SQ16. The multi-touch application passed through an evaluation study?	This question could be answered with yes or no.
SQ17. Experiment description.	This question gathered the summary of the studies presented in each paper.
SQ18. How many participants were there?	This question gathered the number of participants in each study that involved the Usability and/or UX assessment technology.

Table 4. Terms and search string used in SMS.

Population	(“multi-touch interface*” OR “multi-touch recognition” OR “multi-touch-based interaction” OR “multi-touch interaction” OR “multi-touch application” OR “multi-finger interaction” OR “multi-finger interface” OR “multi-finger application” OR “multi-finger recognition”)	AND
Intervention	(“tool” OR “framework” OR “technique” OR “method” OR “model” OR “process” OR “guideline” OR “pattern” OR “metric” OR “approach” OR “inspection” OR “principle” OR “heuristic” OR “methodology” OR “mechanism” OR “questionnaire” OR “checklist”)	AND
Outcome	(“usability evaluation” OR “usability assessment” OR “ux evaluation” OR “ux assessment” OR “user experience evaluation” OR “user experience assessment” OR “user-centred evaluation”)	

Table 5. Number of papers returned and selected.

Data Source	Returned	1st Filter	2nd Filter
ACM Digital Library	185	68	33
Engineering Village	29	18	11
SCOPUS	408	69	21
Total	**622**	**155**	**65**

Some papers appeared more than once in different data sources. They were considered only in the first data source returned, according to the search sequence: ACM, Engineering Village, and SCOPUS, respectively.

3.2.2 Data Extraction.

The extraction of data from our SMS was based on the answers of each SQ. In addition to responses from SQs, we analyzed publications venues (conferences and journals) and years. The technical report containing extractions, graphics and other information is available in FigShare.

3.2.3 Data Analysis.

Researchers transcribed all information obtained in extractions to a document in Microsoft Excel, which helped create the graphs shown in the findings section.

3.3 Phase 3: Reporting

3.3.1 Publication Years.

Extracted papers were published between the years 2008 and 2021. The first study was registered 14 years ago, with its peak in 2013. Figure 4 illustrates the temporal view of the analyzed papers. We believe that the peak noted in 2013 was due to the popularization of smartphones and the news that multi-touch technology provided.

3.3.2 Publication Venues

Only papers approved by a peer-review process were selected. A total amount of 25 different conferences were identified. The conference with most papers analyzed was *ACM Conference on Human Factors in Computing Systems* (CHI), with 12 papers. We analyzed publications from 14 different journals returned in the search, being the *International Journal of Human-Computer Studies* the most returned journal, with four occurrences. The rest of the conferences and journals is presented on the report.

4 Findings

Our main question, ”What technologies are used to evaluate the Usability and UX of software that uses multi-touch-based interaction?” resulted in 29 different technologies found, with 123 occurrences of their use. They are presented in Fig-

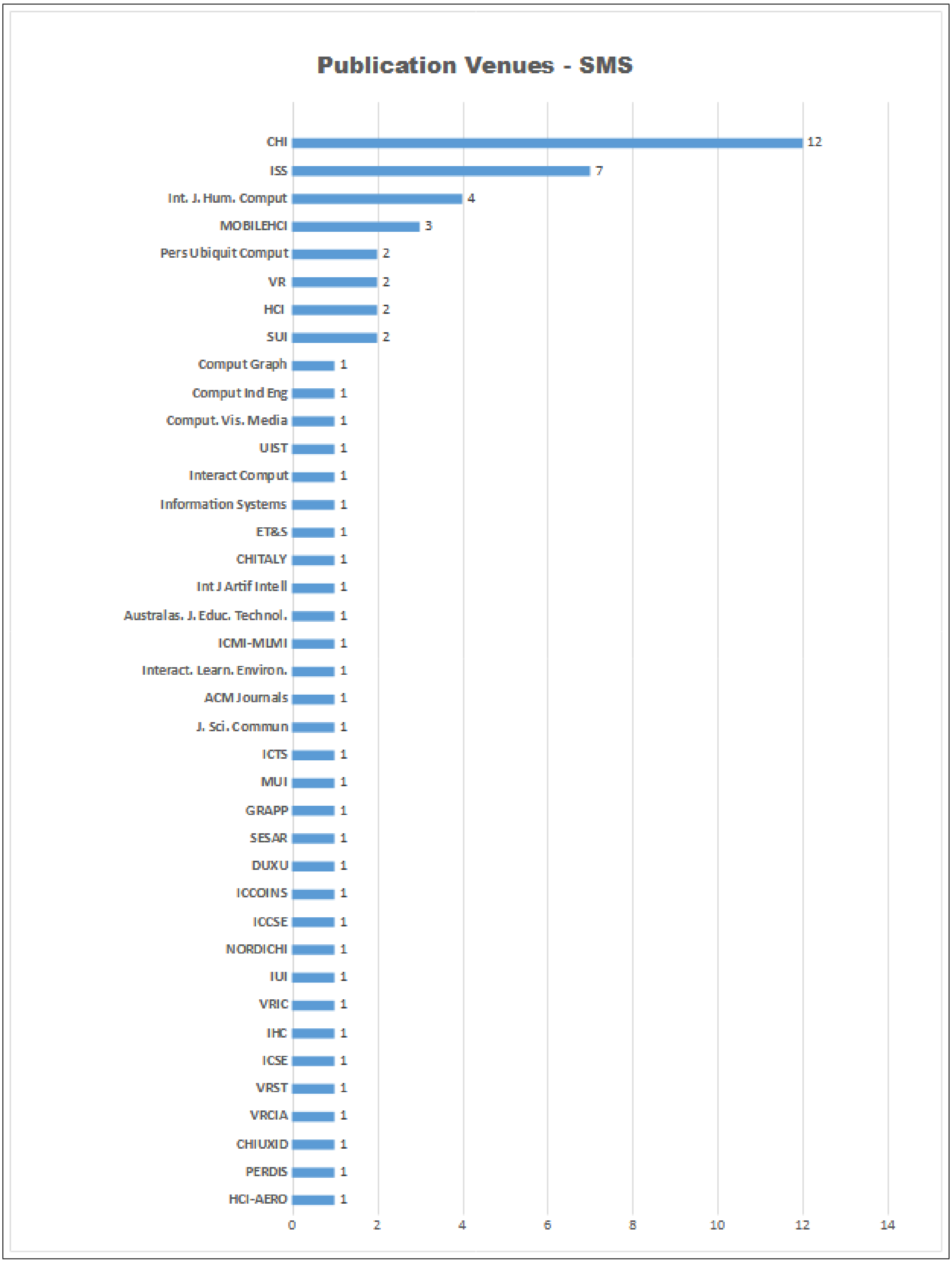

Figure 2. Venues of paper publications found in the SMS

ure 3, with the number of occurrences of each one of them. They were selected and classified based on the authors' descriptions. The most used technology is Usability tests created for the studies, with 31 occurrences, followed by questionnaires created for the studies with 29 occurrences and SUS with 11 occurrences. A summary of these findings can be found in Figure 5. Some SQs were not presented in this table because for having qualitative answers that were too long.

4.1 SQ1. Quality Criterion

Results of SQ1 indicated that 68.29% (n=84) of technologies focused their evaluation on the Usability criterion. For example, in Giesler et al. Giesler et al. (2014), the authors evaluated the precision, speed, and difficulty of the participants should interact with 3D objects through their shadows. The technology used was a questionnaire created for the study, which contained four questions to be answered by a 5-point Likert scale ranging from "very easy" to "very hard".

About 18.7% (n=23) of technologies were used to evaluate the UX of multi-touch systems. In Wang and Lindeman (2012), a questionnaire created for the study evaluated aspects such as fun, motion sickness, and sense of presence. Besides, 13.01% (n=16) of identified technologies were used to evaluate both Usability and UX. Jacucci et al. (2010) used the Measurement, Effects, Conditions Spatial Presence Questionnaire (MEC-SPQ) Vorderer et al. (2004) technology to evaluate experience, concentration, errors, active thinking, and imagination space. The great amount of Usability focused-technologies can indicate that the UX is considered to be less relevant when it comes to evaluate the pragmatic and hedonic set of aspects from the user point of view. Besides, the lack of a jointly evaluation of this concepts can indicate that the most possible complete evaluation is not being done in the majority of the studies.

4.2 SQ2. Usability and/or UX aspects

Results in SQ2 identified 83 different aspects used to evaluate Usability and 76 different aspects to evaluate UX on multi-touch systems. To reduce researchers' bias, we classified the aspects according to the definition and terms provided by the authors of the papers. We observed that some aspects, e.g., efficiency and precision, are considered in some cases as a Usability aspect and in others as UX.

Regarding Usability aspects, we identified the following aspects with the respective number of evaluations: performance (27), overall usability (22), ease of use (20), efficiency (13), effectiveness (11), mental demand (9), frustration (8), effort (8), temporal demand (8), physical demand (8), intuitiveness (7), ease of learning (7), speed (5), accuracy (5), precision (3), understandability (2), utility (2), learneability (2), familiarity (2), quick to use (2), workload (2), satisfaction (2), help (1), remember (1), zoom level (1), number of interactions (1), direct (1), laborious (1), useless (1), useful (1), complicated (1), easy (1), comfort (1), task execution mode (1), mental effort of transition (1), spatial relationship (1), smooth transition (1), usage frequency (1), necessity (1), user acceptance (1), sense of control (1), control (1), concentration (1), goals (1), balance between skill/challenge (1), imagining space (1), activated thinking (1), error (1), ease of sorting (1), ease of selection (1), fun to use (1), tangibility (1), ergonomics (1), reliability of interaction (1), ease of remembering (1), simplicity (1), ease to operate eyes-free (1), ease to perform (1), error recovery (1), difficulty (1), operation mental demand (1), learning performance (1), relevance (1), posture (1), rating of individualization (1), error tolerance (1), conformity of user expectations (1), number of landings (1), overall rating (1), separation violation (1), adoption rate (1), operation smoothness (1), smoothness of transition (1), time to adapt (1), level of integration (1), force (1), concurrency (1), density (1), fragmentation (1), manipulation (1), interaction between floating objects (1), finger and wrist fatigue (1).

The most used aspect of Usability was performance, which was used 27 times. In work presented by Wagner et al. Wagner et al. (2012), the authors calculated performance from the time spent by participants to try, react and perform bimanual tasks on a tablet. Performance probably is the most used aspect due to its intrinsic value, being an aspect that can be used to calculate several quantitative characteristics that provide a good system working overview, and mainly a good non-subjective point, allowing direct comparisons.

Regarding UX aspects, we identified: General UX (5), attraction (4), comfort (3), enjoyment (3), stimulation (3), tension (2), effort (2), preference (2), novelty (2), identity (2), hedonic quality (2), pragmatic quality (2), imagining space (1), presence perception (1), disturbed (1), intuitive control (1), immersion (1), relatedness (1), autonomy (1), competence (1), loss of balance (1), after effects (1), fun (1), nausea (1), motion sickness (1), sense of presence (1), unlikable (1), likeable (1), system composition quality (1), exciting (1), lame (1), clear (1), confusing (1), bad (1), good (1), safety (1), perceived competence (1), pressure (1), negative affect (1), positive affect (1), importance (1), interest (1), concentration (1), ease of use (1), pleasure (1), experience (1), functional disconnection (1), cognitive disconnection (1), perceptual disconnection (1), stimuli comfort (1), strength of stimuli (1), relevance (1), physical interference (1), collaboration sense (1), functionality perception (1), functionality flow (1), participants reaction (1), clarity (1), perspicuity (1), dependability (1), efficiency (1), novelty (1), awareness (1), spatial presence (1), attention (1), motivation (1), precision (1), reality factors (1), distraction (1), sensitivity (1), control factors (1), general feeling (1), intuition (1), verbalization (1), magical experience (1), perceived difficulty (1), fatigue perception (1).

The most used aspect was the overall UX, where the authors did not define any specific aspect and evaluated UX in a generalized way. For example, Leftheriotis (2013) used videos and images to evaluate the user's overall UX when using software to increase security while inserting passwords in public. As the UX can be seen as an hedonic concept, it seems adequate to ask to the user an general overview about the feelings produced by the lived experience.

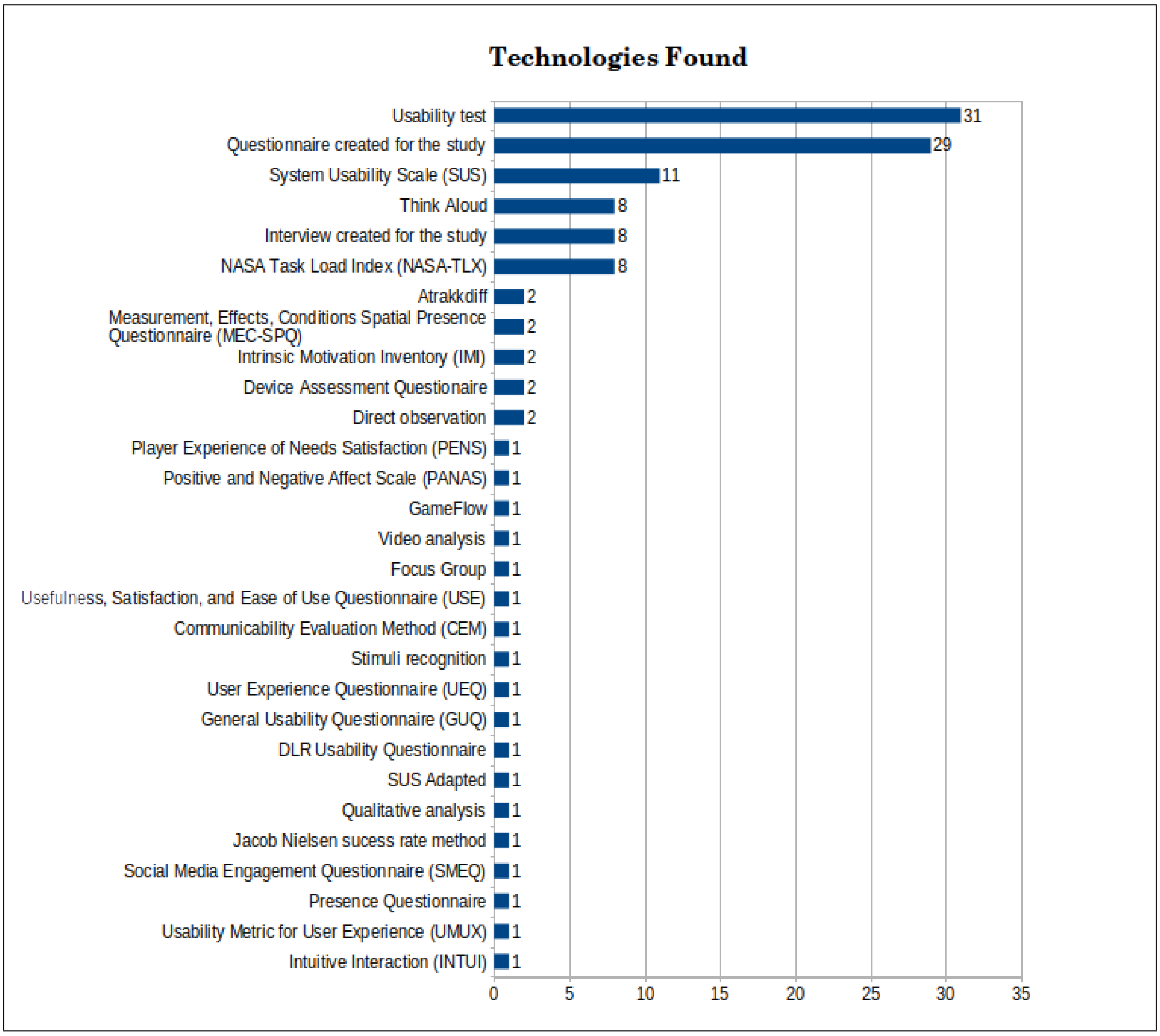

Figure 3. Technologies found in this SMS

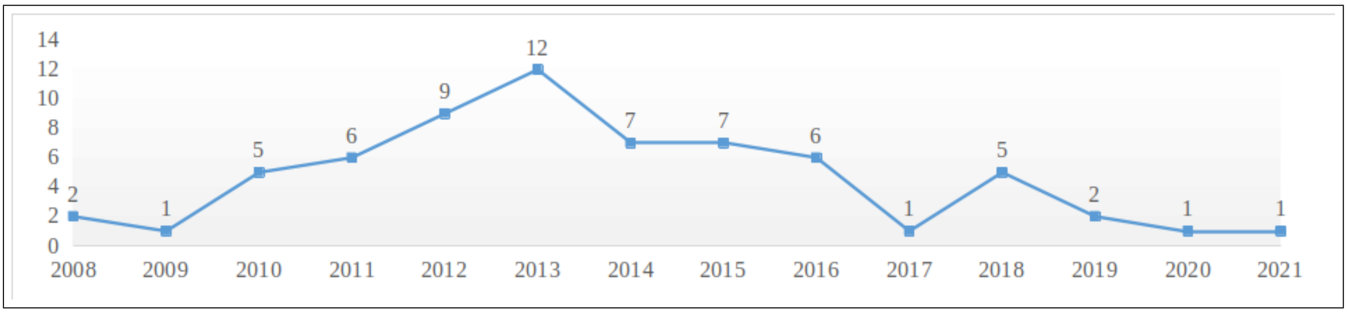

Figure 4. Temporal view of papers analyzed in the SMS.

4.3 SQ3. Specificity of evaluation technology

About 96.75% (n=119) of technologies returned are not specific to multi-touch systems. They can evaluate any software for multi-touch, web systems, and voice interaction, among others. Uebbing-Rumke et al. (2014) used SUS (Brooke, 1996) to evaluate Usability of a air traffic control simulator. SUS is an evaluation technology used in different contexts and contains ten statements about Usability that can be answered by a 5-point Likert scale.

About 3.25% (n=4) of technologies were specific to the multi-touch context. For example, Ghomi et al. (2013) developed a questionnaire with specific questions about an educational system developed to help students to understand and manage money, coins, and banknotes. This experiment lasted three weeks, having collaborative work, and by the end of the period, the students answered the questionnaire that focused on motivation. This results show that the necessity to elaborate a system taking into account the multi-touch context specificities is not perceived by the majority of studies authors. This lack of specification can lead to not perceive some main problems intrinsic to the multi-touch context, as prejudiced UX due to lack of touch precision (Forlines et al., 2007), lack of mechanical feedback (Buxton et al., 2007) and lack of gesture standardization (Liang et al., 2011).

4.4 SQ4. Basis of evaluation technology

Results of SQ4 revealed 35.77% (n=44) of technologies used are based on an existing one. For example, in Watson et al. (2013), the authors used PENS technology (Ryan et al., 2006) to evaluate the Usability of a game. PENS assesses competence, autonomy, relatedness, immersion, and intuitive control through a 5-point Likert scale.

In contrast, 64.23% (n=79) of technologies was explicitly created for the study reported in the paper. For example, in Kim et al. (2012), the authors created a questionnaire to evaluate a system developed for photo triaging. This questionnaire contained questions about Usability that were answered by a 5-point Likert scale. The higher number of created technologies can indicate that a lot of studies need to evaluate specific contexts, where cannot be found existing technologies adequated for the context.

4.5 SQ5. Method of data collection

SQ5 showed that 20 methods were identified to collect data from participants. We verified the following methods that are presented in Figure 7. The most used collection method is the Likert scale, used on 62 occasions, where users answered an agreement scale according to the statement provided. For example, Tseng et al. (2018) used a Likert scale to collect responses from 6 statements that were answered in 5-points. Also, we verified other methods used several times, such as time spent on tasks, system log files, and open-ended questions (qualitative analysis of users' answers). In 8 papers, the collection method was unclear.

Table 6. References found in SMS.

Zhang et al. (2020)	Gürlük et al. (2014)	Giesler et al. (2014)
Colley et al. (2015)	Kim et al. (2012)	Pfeiffer et al. (2010)
Jacucci et al. (2010)	Fu et al. (2010)	Liang et al. (2011)
Damaraju et al. (2013)	Wang and Lindeman (2014)	Wingert et al. (2017)
Jetter et al. (2011)	Ackad et al. (2010)	Wang and Lindeman (2012)
de Souza Alcantara et al. (2012)	Huerta et al. (2011)	Watson et al. (2013)
Mossel et al. (2013)	Bertolo et al. (2013)	Echtler et al. (2009)
Ghomi et al. (2013)	Olwal et al. (2008)	Rädle et al. (2013)
Leftheriotis (2013)	Micire et al. (2011)	Wagner et al. (2012)
Freitag et al. (2012)	Coram et al. (2013)	Frisch et al. (2011)
Telkenaroglu and Capin (2013)	Lissermann et al. (2014)	Tuddenham et al. (2010)
Chen et al. (2016)	Lammel et al. (2016)	Nacher and Jaen (2015)

4.6 SQ6. Characteristics of evaluation technologies

SQ6 was answered with open responses, varying according to the technology. About the most used technology (Usability Tests), SQ6 extracted the methods used and the data collected. For example, Nacher and Jaen (2015) registered the percentage of the number of repetitions performed successfully, error rate, and completion time in a log file to measure the accuracy of pre-kindergarten children.

About the second most used evaluation technology, questionnaires created for the study, SQ6 extracted all questions contained in the questionnaire. For example, in Echtler et al. (2009), authors created a questionnaire with six statements to evaluate enjoyment, disturbance, perception of presence, and efficiency of a Sudoku game running on a public display and mobile devices carried by passing users. Among statements, we can cite "*Have you been disturbed by the actions of other players when you played on the handheld?*" and "*How present were the other players when you played at the tabletop?*".

SQ6 extracted the characteristics of each evaluation technology. From questionnaire and interview technologies, questions and statements were extracted. Methods and data were extracted from analysis technologies. From metric and model technologies, the calculation was extracted. Technologies details were not covered in our paper due to space limitations.

Research SQ	Possible Answers	Findings: Technologies	Findings: Percentage %
SQ1. What is the quality criterion of evaluation technology used?	Usability	84	68.29
	User eXperience	23	18.7
	Both	16	13.01
SQ3. Is the technology specific to multi-touch systems or to systems in general?	Specific	4	3.25
	Generic	119	96.75
SQ4. Was the technology created for the study or is it based on an existing one?	Existing	44	35.77
	Created	79	64.23
SQ7. Does technology extract quantitative or qualitative data?	Quantitative	98	79.67
	Qualitative	23	18.7
	Both	2	1.63

Research SQ	Possible Answers	Findings: Occurrences	Findings: Percentage %
SQ10. Was the software created for a specific group of people? Which groups?	Yes	14	21.21
	No	52	78.79
SQ11. How many fingers are used to interact with the multi-touch system?	2	53	50.96
	3	14	13.46
	4	9	8.65
	5	5	4.81
	6	2	1.92
	7	2	1.92
	8	2	1.92
	9	2	1.92
	10	3	2.88
	More than 2	12	11.54
SQ12. What is the gesture used in the multi-touch interaction?	Drag	37	33.94
	Pinch	31	28.44
	Rotation	18	16.51
	Touch	11	10.09
	Touch and Drag	4	3.67
	Free Gesture	1	0.92
	Undefined	7	6.42

Figure 5. Summary of findings

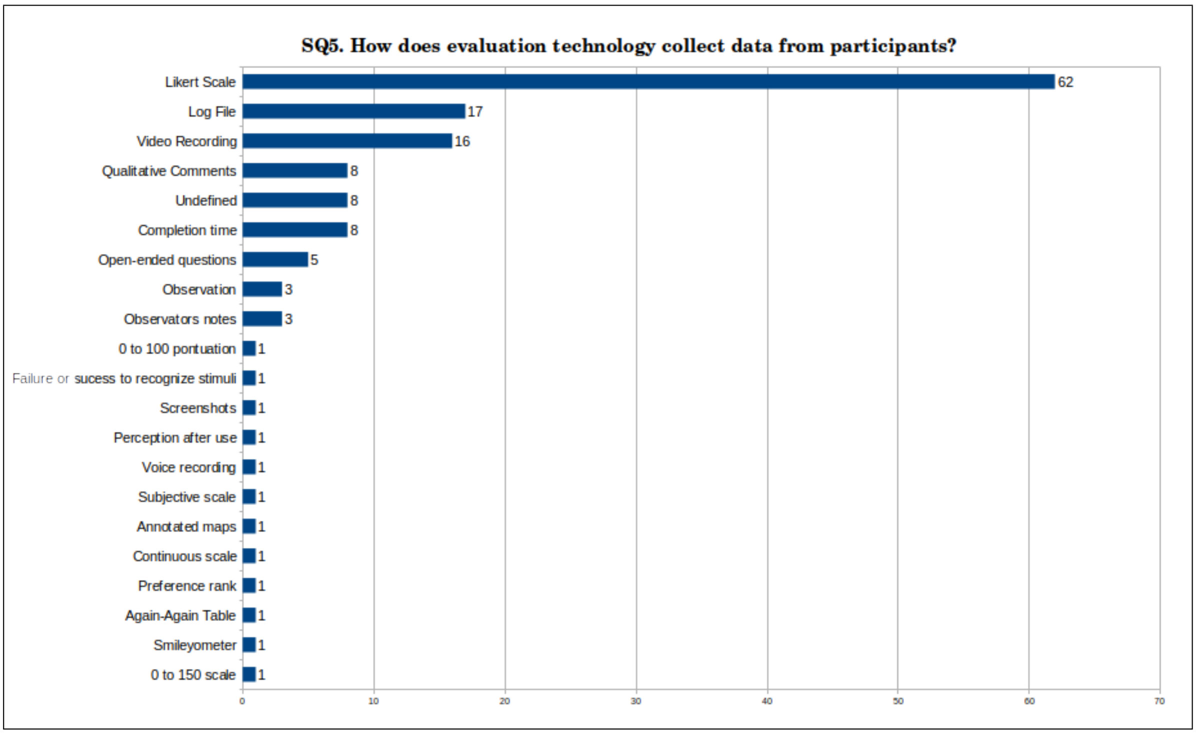

Figure 6. Twenty different methods for data collection were found

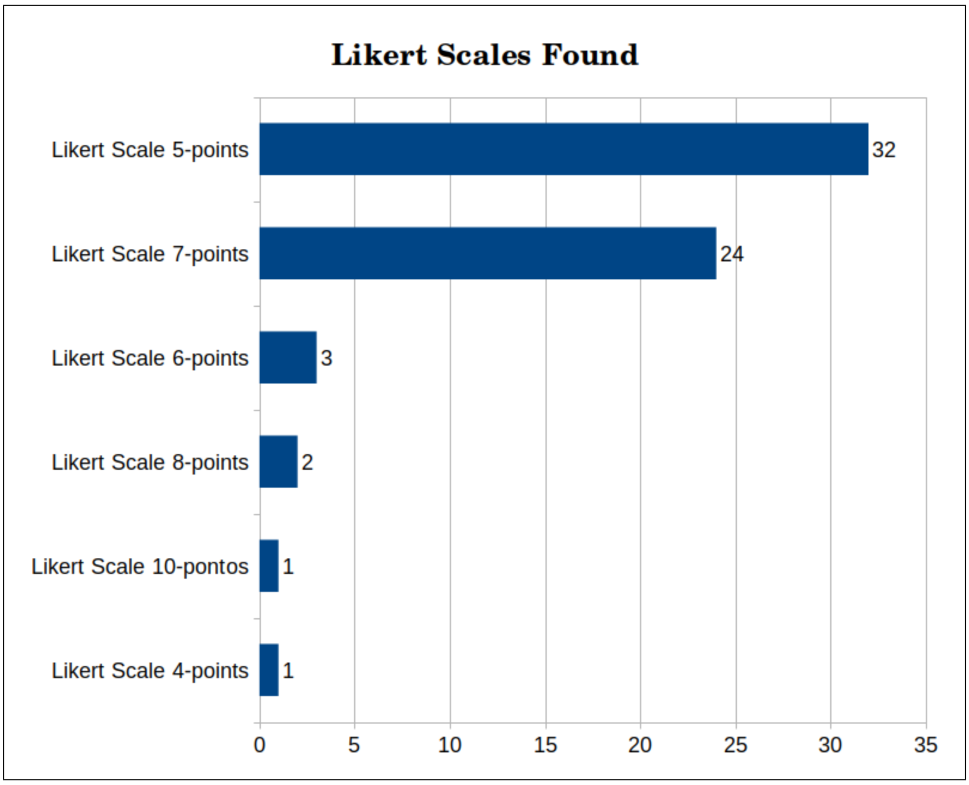

Figure 7. Distribution of Likert Scales found in the technologies

4.7 SQ7. Analysis type

Results of SQ7 revealed that 79.67% (n=98) of technologies evaluated the multi-touch systems quantitatively. Martin-SanJose et al. (2017) performed a quantitative analysis based on responses obtained with a 5-point Likert scale. The scale ranged from ”never” to ”always”, which answers fourteen questions regarding motivation.

Approximately 18.7% (n=23) of the technologies evaluated data qualitatively. Renzi and Freitas (2014) analyzed the study's voice and video recording to perform a qualitative analysis based on insights obtained with the Think Aloud method. Think Aloud aims to identify possible thoughts from users based on what they say loud during the experiment.

Only 1.63% (n=2) of technologies evaluated data quantitatively and qualitatively. Freitag et al. (2012), authors created a questionnaire that contained questions to be answered with a 5-point Likert scale, ranging from ”low” to ”high”, with space for open comments.

It is worth of note this so low percentage of technologies jointly evaluating the quantitative and qualitative data. It is valid to assume that this significant amount of one-focused aspect technologies may be due to less effort required to evaluate only essential research data. After all, not every study needs to extract and evaluate both data types. However, the tiny proportion of 1.63% of technologies that evaluate both aspects together is enough to ask whether the potential of this evaluation is not being overlooked.

4.8 SQ8. Multi-Touch System function

SQ8 extracted subjective data about multi-touch systems and was used to capture a functionalities description of the evaluated software. For example, in Coram et al. (2013), the authors evaluated a system for analysis and visualization in astrodynamics, supporting space mission analysis. Besides, Fabroyir (2019) evaluated a map system, where the interaction through multi-touch was compared with the interaction of game control.

4.9 SQ9. Multi-Touch System category

SQ9 presents a categorization of systems identified in SQ8. The categories with the number of papers were: daily facilites (24), virtual reality (11), amusement (7), mobility (3), security (3), education (2), health (2), augmented reality (2), image edition (1), aeronautical mobility (1), music (1), 3D interaction and simulation (1), criativity (1), accessibility (1), spacial visualization (1), 3D navigation environment (1), object manipulation (1), social networks (1). In one case the category could not be defined.

Leftheriotis et al. (2015) addressed daily facilities, where they evaluated a system made to identify plant species. About health systems, Madni et al. (2016) evaluated two different orientation techniques for diagnosing and monitoring medical images from the user perspective. Fu et al. (2010) evaluated an educational system made for the exploration of 3D astrophysics simulations.

4.10 SQ10. User group of Multi-Touch System

Results of SQ10 revealed that 78.79% (n=52) of evaluated systems were not created for a people group. In contrast, 21.21% (n=14) of systems were created to help some groups. Groups and the number of multi-touch systems created to target them are Drivers (2); Students (2); Blind/Visually

Table 7. References found in SMS

Madni et al. (2016)	Renzi and Freitas (2014)	Derboven et al. (2012)
Uebbing-Rumke et al. (2014)	Rodriguez-Conde and Campos (2020)	Tseng et al. (2018)
Kim and Lee (2015)	Wang and Lindeman (2015)	Merrad et al. (2022)
Fabroyir (2019)	Muender et al. (2019)	Kulik et al. (2018)
Ducasse et al. (2018)	Cascales-Martínez et al. (2016)	Arnaud et al. (2016)
Leftheriotis et al. (2015)	Schürmann et al. (2015)	Nacher et al. (2015)
Hsiao et al. (2014)	Bertolo et al. (2013)	Kildal et al. (2013)
Blažica et al. (2013)	Koutlemanis et al. (2013)	Tuveri et al. (2013)
Zaharias et al. (2013)	Radhakrishnan et al. (2013)	Ciocca et al. (2012)
Berkman and Karahoca (2012)	Hachet et al. (2011)	

Impaired (1); Air Traffic Controllers (1); Design Experts (1); Medics (1); Scientists with specialties in biochemistry (1); Astrodynamics Specialists (1); Aeronautics Professionals (1); Editors (1); Children (1); Fashion Designers (1).

The drivers and students were the groups with the greatest targeting of returned multi-touch systems. For example, in Colley et al. (2015), the authors evaluated the Usability and UX of a multimedia system of a car.

4.11 SQ11. Number of fingers used

SQ11 searched to find the number of fingers used in the multi-touch systems. The number of fingers and their occurrences are: 2 fingers (53), 3 fingers (14), 4 fingers (9), 5 fingers (5), 6 fingers (2), 7 fingers (2), 8 fingers (2), 9 fingers (2), 10 fingers (3). In 12 papers, we could not identify the exact number of fingers used, but for sure, there were more than 2. For example, in Micire et al. (2011), the authors evaluated an interface that used 10 fingers to pilot a robot. The greater set of 2 fingers makes sense when thinking that it can make it easier to perform gestures with fewer fingers. The most used daily systems also use 2 fingers to perform their main multi-touch actions, like Google Maps and Google Earth.

4.12 SQ12. Gesture used

In SQ12, we identify and classify the gestures used in the studies. Due to the lack of a previous classification, we divided the gestures into generic categories that could comprise the varieties of gestures found. The categories and the occurrences they were used are: Drag (37), Pinch (31), Rotation (18), Touch (11), Touch-and-Drag (4), and Free Gesture (1). In 7 cases, we could not identify the gestures. For example, in Pfeiffer et al. (2010), the participants were asked to create 19 different gestures to interact with a music player and a navigation system in a car.

4.13 SQ13-SQ15. Evaluation study

of evaluation technology

About SQ13, we verified that if the evaluation technologies passed through an evaluation study by the authors, what is an important step to check the limits and quality of the technology. Nonetheless, none of the technologies found were empirically evaluated, even when they were created for the study, in the case where this validation would be more crucial (Shull et al., 2001). Regarding SQ14 and SQ15, we also got no results since there was no evaluation study.

4.14 SQ16-SQ18. Evaluation study

of multi-touch technology

SQ16 verified if there was or was not an evaluation study of the multi-touch interface presented in the papers. Through SQ16, we could verify that all multi-touch-based interfaces passed through an evaluation study. Then SQ17 summarized the experiment conducted in each paper to facilitate the understanding, analysis, and future work based on each extraction. Finally, SQ18 collected the number of participants present in each study. The amount of participants is very varied, ranging from 2 to 326. Rodriguez-Conde and Campos (2020) performed a case study with 326 participants to find out the differences between interaction through multi-touch and a desktop.

5 Benchmark

Due to the several technologies found in the SMS, a benchmark was performed to characterize better and analyze the questionnaires identified by the SMS. A benchmark is a method used to measure, compare, define best practices, implement, and improve a software or product (Anand and Kodali, 2008). The General Usability Questionnaire (Uebbing-Rumke et al., 2014) and the DLR Usability Questionnaire (Uebbing-Rumke et al., 2014) were not included in this analysis because the authors do not provide them. For the same reason, the questionnaires created specifically for the studies where they were applied were not selected, and they can be seen in Table 8.

A total of 14 questionnaires were cataloged and the number of questions was also considered an aspect of to catalog (Tables 9 and 11). The criterion targeted by the technologies were classified as Usability, UX, or both. Observing the questions, two different types were noted: extensive questions that can formulate a line of thinking and just aspects that opposed each other in different extremes of scales. The scales, in turn, also became an analyzed aspect since their presence was perceived in all questionnaires. The number of aspects evaluated by the questionnaires also was counted, both in number and in criteria, this criterion being Usability or UX. We also checked whether the questionnaires had automated versions that could facilitate collecting and analyzing the answers.

Table 8. Authoral questionnaires found in SMS

Colley et al. (2015)	Giesler et al. (2014)	Kim et al. (2012)
Jetter et al. (2011)	Zhang et al. (2020)	Wang and Lindeman (2014)
Ackad et al. (2010)	Wang and Lindeman (2012)	Liang et al. (2011)
Huerta et al. (2011)	Echtler et al. (2009)	Ghomi et al. (2013)
Bertolo et al. (2013)	Freitag et al. (2012)	Mossel et al. (2013)
Lissermann et al. (2014)	Telkenaroglu and Capin (2013)	Tuddenham et al. (2010)
Olwal et al. (2008)	Chen et al. (2016)	Kim and Lee (2015)
Wang and Lindeman (2015)	Tseng et al. (2018)	Cascales-Martínez et al. (2016)
Hsiao et al. (2014)	Bertolo et al. (2013)	Ducasse et al. (2018)
Zaharias et al. (2013)	Hachet et al. (2011)	Hachet et al. (2011)

The number of questions found ranges from four (PENS and UMUX) to 45 (IMI). The number of questions implies how much information can be provided by the user. Questionnaires that want to extract much information should consider that more questions can be beneficial. However, when the questionnaire's objective is a quick answer and not too deep about what they need to discover, a few questions can help make the process smoother and faster. The amount of questions also implies how many usability and/or UX aspects can be assessed.

The predominant criterion targeted by the questionnaires is UX, being 57.14% of the sample (n = 8), followed by usability with 35.71% (n = 5). Only MEC-SPQ seeks to evaluate usability and UX jointly, representing 7.14% (n = 1) of the analyzed group. The joint evaluation of usability and UX allows verification, simultaneously, aspects linked to behavioral goals and those linked to users' feelings (Väätäjä et al., 2009). The evaluation of UX and usability in separate ways, predominant in this set of questionnaires, may indicate that a complete analysis of user behavior is usually performed using more than one evaluation technology.

Regarding the scales used, 57.14% (n = 8) use 7-point scales: Attrakdiff, IMI, INTUI, NASA-TLX, Presence Questionnaire, UEQ, UMUX, and USE. The questionnaires that use 5-point scales are DAQ, MEC-SPQ, PANAS, and SUS, corresponding to 28.57% (n = 4) of the sample. SMEQ uses an 8-point scale, which is 7.14% (n = 1) of the sample. SMEQ evaluates how many days of the week the user utilizes social media. Since the week has seven days, the extra point corresponds to the option "I do not use social media". PENS uses a variable scale of 2 to 3 points, being 7.14% (n = 1) of the occurrences. PENS has this variable scale because some of their questions give the option of rating one or two stars, while others give a rating of one to three.

Discussing the number of points used in an evaluation technology can guide a researcher to use an optimized number of questions for the context. Chyung et al. (2017) make an analysis that shows that the most important reason when choosing the number of points is the presence of a middle point. A middle point in a scale can allow the user to choose a minimally acceptable response as soon as it is found instead of putting effort into finding an optimal response (Matell and Jacoby, 1971). The name "satisfying behavior" is given to this phenomenon. In this way, when the researchers have to minimize incorrect middle point use, it is advisable to omit this point or increase the scale sensitivity. There are some other situations where the middle point omission is recommended. They are: when respondents are comfortable with the survey topic, offering an "I do not know" option instead; when they are under intense pressures of social desire (Johns, 2005); and when they have little or no involvement with the topic (Weems and Onwuegbuzie, 2001). However, there are situations where the researchers can have less biased results on applying the middle point on the scale. On vague topics, a neutral opinion can be desired (Johns, 2005); a neutral point can improve instrumental confidence when measuring psychological characteristics (Adelson and McCoach, 2010); some measures can be taken to decrease the erroneous use of the neutral point by improving the clarity of the questionnaire items (Kulas and Stachowski, 2013).

The type of scale used varies. Around 42.86% (n = 6) use a semantic differential scale, where there is an opposition between the meanings of each end of the scale. They are AttrakDiff, DAQ, IMI, INTUI, NASA-TLX, and UEQ. On the other hand, in simple Likert scales, where each extreme of the scale indicates total agreement or disagreement with a term, we have 50% of use (n = 7), being them MEC-SPQ, PANAS, PENS, Presence Questionnaire, SUS, UMUX, and USE. In just one case (n = 1), representing 7.14% of occurrences, we have SMEQ that used the days of the week as a measure, not fitting the above mentioned standards. Using questions or just opposing aspects can imply advantages and disadvantages. Extensive questions can explain better for the participant the meaning of the question, leading him to give a more accurate answer. However, if the researcher desires to approach many questions, it could tire the user, leading him to pick a point to finish it faster. From this point of view, using just opposing aspects in extensive questionnaires could be an advantage. These thoughts lead to a belief that the researcher must choose a format that better suits the context in which the questionnaire will be applied.

The number of aspects assessed by questionnaires varies greatly, with the highest being 56 and the smallest being 1. The questionnaire that evaluates only one aspect is SMEQ, with the frequency of social media use being the target. All questionnaires that use a semantic differential scale present two aspects, one on each end of the scale; therefore, the number of questions is usually half the number of aspects evaluated. This approach is based on the perception that the concepts and their antonyms cannot be combined into a single concept to apply to each question. This approach occurs because it was noted that different authors could interpret some concepts differently. It is interesting to note that 5 of the 14 questionnaires evaluate four different aspects, the largest group with an equal number of aspects of this benchmark. However, only UMUX has four questions, one for each as-

Table 9. Result of the benchmark

Questionnaires	Number of Questions	Criteria	Scales	Questions for each aspect
AttrakDiff	28	UX	7 points	No
DAQ	13	Usability	5 points	Yes
IMI	45	UX	7 points	Yes
INTUI	17	UX	7 points	Yes
MEC-SPQ	20	Both	5 points	Yes
NASA-TLX	6	Usability	7 points	Yes
PANAS	20	UX	5 points	No
PENS	4	UX	2-3 points	No
Presence Questionnaire	29	UX	7 points	Yes
SMEQ	5	UX	8 points	Yes
SUS	10	Usability	5 points	Yes
UEQ	26	UX	7 points	No
UMUX	4	Usability	7 points	Yes
USE	30	Usability	7 points	Yes

Table 10. Result of the benchmark

Questionnaires	Collection Methods	Aspects Evaluated	Automated version?
AttrakDiff	Differential Semantic Scale	28-56	Yes
DAQ	Differential Semantic Scale	9	No
IMI	Differential Semantic Scale	11	No
INTUI	Differential Semantic Scale	4	No
MEC-SPQ	Likert Scale	7	No
NASA-TLX	Differential Semantic Scale	6	Yes
PANAS	Likert Scale	20	Yes
PENS	Likert Scale	5	No
Presence Questionnaire	Likert Scale	4	No
SMEQ	Amount of days using social media per week	1	No
SUS	Likert Scale	4	Yes
UEQ	Differential Semantic Scale	26-52	Yes
UMUX	Likert Scale	4	No
USE	Likert Scale	4	Yes

Table 11. Questionnaires catalogued in the benchmark

Attrakdiff (Hassenzahl, 2004)	Device Assessment Questionaire (DAQ) (ISO 9241-9, 2000)	Instrinsic Motivation Inventory (IMI) (Ryan and Deci, 2000)
Intuitive Interaction (INTUI) (Ullrich and Diefenbach, 2010)	Measurement, Effects and Conditions – Spatial Presence Questionnaire (MEC-SPQ) (Vorderer et al., 2004)	NASA-Task Load Index (NASA-TLX) (Hart, 1986)
Positive And Negative Affect Scale (PANAS) (Watson and Clark, 1994)	Player Experience of Need Satisfaction (PENS) (Ryan et al., 2006)	Presence Questionnaire (Witmer and Singer, 1998)
Social Media Engagement Questionnaire (SMEQ) (Przybylski et al., 2013)	System Usability Scale (SUS) (Brooke, 1996)	User Experience Questionnaire (UEQ) (Laugwitz et al., 2008)
Usability Metrics for User Experience (UMUX) (Finstad, 2010)	Usefulness, Satisfactions and Ease of Use Questionnaire (USE) (Lund, 2001)	

pect. The others have 10 to 30 questions, divided into four groups, each focusing on the aspect they want to evaluate. This observation leads us to believe that it is essential when seeking to fully evaluate an aspect, to formulate more than one question about it.

From the 14 questionnaires, 42.86% (n = 6) have an online version. They are: Attrakdiff [6], NASA-TLX[7], PANAS[8], SUS[9], UEQ[10] and USE[11].

Attrakdiff has an online tool available in English and German. It offers two versions, a complete one and a shorter one. It meets some target profiles of the evaluation, namely the individual evaluation, comparison between A and B, and comparison between before and after. Its primary focus is on evaluating the attractiveness of products. Creating the online form allows for customizing various information, such as the product's name and the type of product being evaluated, such as medical devices or logistical systems. It allows the creation of the form to be distributed anonymously, or if the evaluator has the participants' emails, it allows the invitation to be sent directly by email. At this point of configuration, the tool gives you information about the created project, the possibility of modifying the date so that the form expires at some point, and information about the participants' test. The possibility of exporting the results as an Excel table is given at the end of the results section. It also generates some graphs to evaluate the participants' behavior that can be easily exported. We made a test, and the exported images can be seen in Figure 8. This online tool is easy to apply; it is intuitive and can be very helpful in evaluating the proposed aspects.

NASA-TLX automated version is officially available on the USA govern site[12]. It proposes to evaluate six aspects in order to be able to measure the workload of a given activity. Its automated availability is limited to Apple devices such as iPhone, iPad, iPod, and Mac. The app presents a series of 21 steps, the first 15 pairwise systems, wherein each pair, the user must choose the more relevant aspect regarding workload in the task recently performed. After that, the six main questions in the paper and pencil versions are presented. That version uses a 7-point Likert scale, but the automated version gives a range of choices that is way more accurate, as seen in Figure 9. After this, the user can access the app history. Each new evaluation generates two graphics, one for the pairwise evaluation and one for the rating scale, as seen in Figure 10. For the pairwise approach, weights are assigned to the six evaluated aspects. However, the app does not clarify how this weighting is done. As for the rating scale approach, the app presents data on weights, rating, an adjustment combining weights multiplied by ratings, and a general weighted rating. The use of the app seems to have the objective of a more personal approach since it does not provide the possibility to collect data and evaluate how several people evaluate a task after its completion. There were no options for sharing the

[6]https://www.attrakdiff.de/index-en.htmltab-einzelausw
[7]https://nasa-tlx.firebaseapp.com
[8]https://psytests.org/emo/panasen.html
[9]https://uiuxtrend.com/sus-calculator/
[10]https://www.ueq-online.org
[11]https://garyperlman.com/quest/quest.cgi?form=USE
[12]https://humansystems.arc.nasa.gov/groups/tlx/

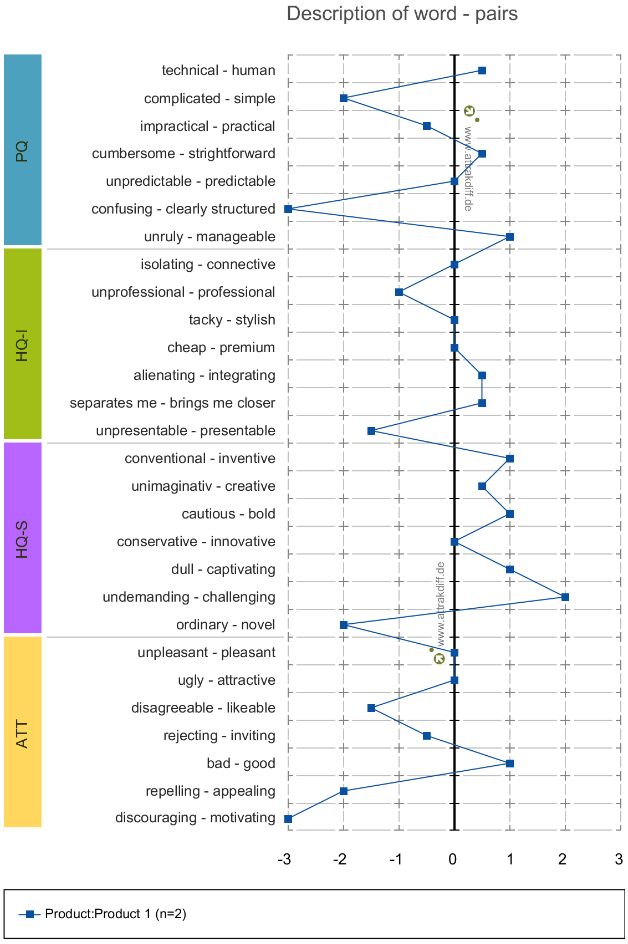

Figure 8. The mean values of the word pairs. Of particular interest are the extreme values, showing which characteristics are particularly critical or particularly well-resolved.

results, which could yield a new version for broader use of this tool.

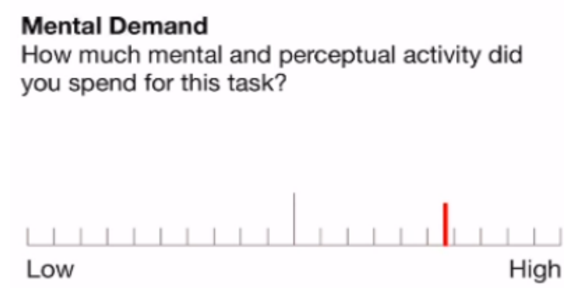

Figure 9. The red pin shows that there is a very wide point picking accuracy.

As for PANAS, two online versions were found. No information was found to confirm if some of the online tools were made by the authors of the original questionnaire. PANAS seeks to measure the positive and negative effects felt last week. The researcher can orientate the object of this affection. The version provided by psytests.org[13] provides some calculations on the 20 questions that can be seen in Figure 11. The site is described as follows: "PsyTests.org is a free-to-use website dedicated to providing psychology professionals, students, and the general public with access to academically validated psychological assessment tools with a simple

[13]https://psytests.org/emo/panasen.html

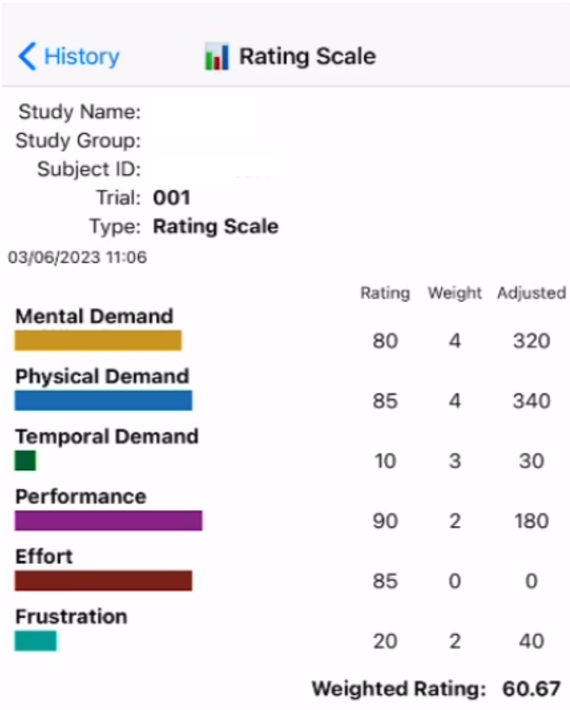

Figure 10. Graphic generated after the rating scale evaluation.

interface and automated scoring". The other online version found was provided by the site novopsych.com.au/[14] and the description of the site is "NovoPsych's mission is to help mental health services use psychometric science to improve client outcomes". For the results, this version can only receive by email. The document generated provides the basic patient information, besides displaying a graphic about the positive and negative affects that can be seen in Figure 12 and giving the instructions to interpret it. Both versions have the same questions and scales. It can be noticed in both versions that they were made for one-on-one contact between patients and mental health professionals. This perception is due to the perceived intention of the questionnaires to be individual assessments. None of the tools allow the formation of calculations or graphs with information from more than one individual. Therefore, a suggestion for future work would be implementing a tool capable of evaluating the results of groups of individuals to understand how these groups feel and perceive the aspects covered by the questionnaire.

SUS is one of the most known questionnaires for evaluating usability. Surprisingly, no free automated online version of deployment, distribution, and results gathering was found. The .gov site[15] presents a template in PDF that can be downloaded. For the result analysis, we found many tools, the most remarkable presented by Blattgerste et al. (2022)[16]. It is an open-source project proposing a SUS analysis toolkit. It allows input data to be analyzed manually or through CSV files. The options to make a single or multiple variable analysis is presented too. As a result, it presents four options of presets, each generating different varieties of charts. One of them is presented in Figure 13. The analysis made by this project is complete, and the integration with an automated way to distribute the questionnaire and gather its answers could result in a more complete and definitive online version of SUS.

UEQ is one of the most known questionnaires regarding UX evaluation. It attends more than 30 languages and is provided in the official site[17]. The project's creators also provide shortened versions of the UEQ and very complete and automated Excel spreadsheets to generate the most diverse graphical and statistical results. The same suggestion to the tools found for SUS can be made for UEQ. Here is a lack of automated ways to collect the participants' response to the questionnaires and apply it directly to the analysis tools.

Only one version of USE[18] was found. It is an essential tool that contains the questions and options to mark them and send the answers by email. In the test made for this benchmark, the answers went to the spam inbox. A feature of adding a comment on any of the items is presented. The aspects evaluated by USE (Usefulness, Satisfaction, and Ease of use) are recurrent in the literature analyzed. Thus, other technologies that already assess these aspects may do so differently or better or be more widespread. Another possibility is that this questionnaire was not well-known and widespread until then.

6 Discussion

Our analysis through this SMS on technologies used to evaluate Usability and/or UX in multi-touch systems revealed 29 different technologies. Furthermore, these technologies had 123 total occurrences when accounting for repeated technologies.

A lack of a standard in the used technologies was identified, with authors preferring to create their evaluation technologies. This identification is evidenced by the high number of Usability tests and questionnaires created for the studies. However, while creating your technology allows you to address the specific issues of the project, it makes it difficult for the technology to be reused by others. In addition, all technologies created did not undergo an empirical evaluation, which can threaten their validity.

6.1 SMS discussion

About SQ1, the small number of technologies that evaluate Usability and UX together is remarkable, as can be seen in Guerino and Valentim (2020), where from 110 technologies found that evaluate Usability and UX on Natural User Interfaces (NUIs), only 2 evaluate them jointly. Most technologies assess Usability aspects. It is possible that with the development and use of technologies that focus on both, a greater understanding of software quality will be achieved.

Regarding SQ2, it is evident that the concepts of UX and Usability are not the same among the authors since aspects such as comfort and effort are considered UX by some authors and Usability by others. Following the aspects that compose Usability according to the ISO/IEC 25010 (2011), we identified that two of the three main aspects are present in

[14] https://novopsych.com.au/assessments/formulation/positive-and-negative-affect-schedule-panas/
[15] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html
[16] https://mixality.de/sus-analysis-toolkit/
[17] https://www.ueq-online.org
[18] https://garyperlman.com/quest/quest.cgi?form=USE

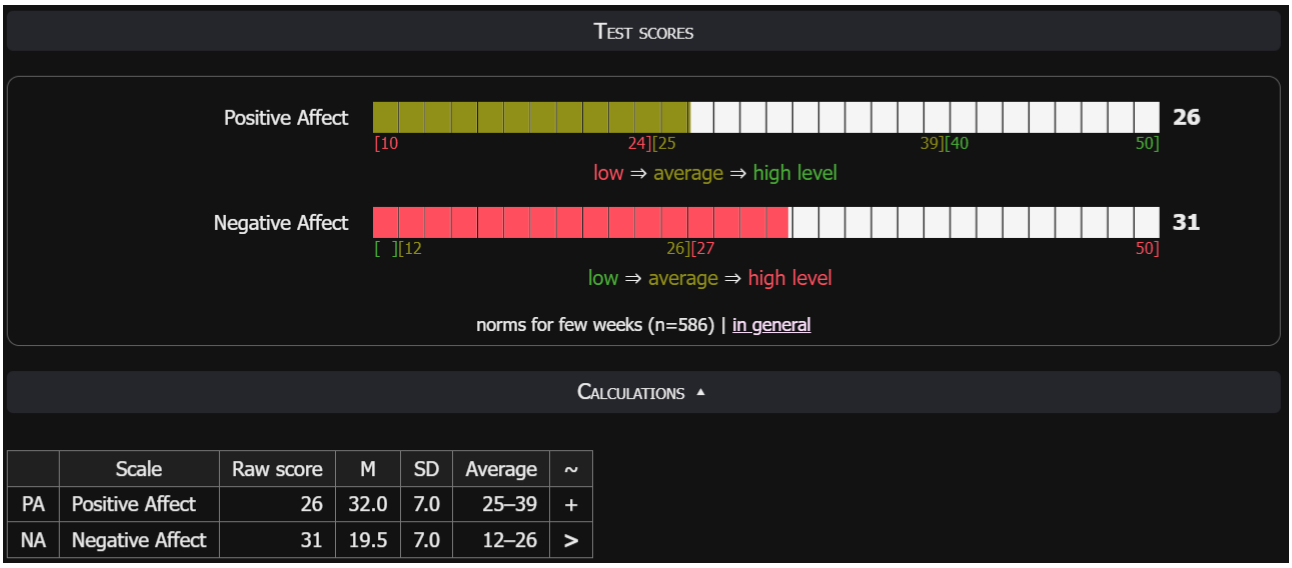

Figure 11. Metrics and calculations made upon the results of PANAS provided by PsyTests.

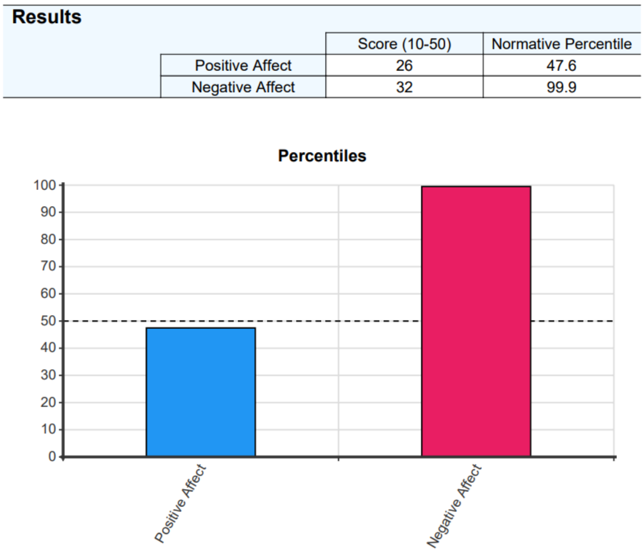

Figure 12. Scores and graphic made upon the results of PANAS provided by NovoPsych.

Usability evaluations, which are effectiveness and efficiency. About UX, authors prefer to evaluate the general user experience over other specific aspects.

In SQ3, a discrepancy is found between technologies specific to evaluating multi-touch interactions and generic technologies. Although generic technologies have the advantage of being used in other studies, we believe that their generic way of evaluating does not fulfill its role well in the multi-touch context. For a better understanding and evaluation, this context needs ways to assess the specificities of multi-touch, i.e., the number of fingers and gestures used. The lack of these specific ways to evaluate this context can negatively impact the evaluation process and the quality of these systems. This negative impact can occur in different ways. Forlines et al. (2007) demonstrates that the lack of precision when detecting touches or gestures is a point that can significantly harm the user experience. Similarly, the lack of mechanical feedback is something that should be taken into account, due to the ability to generate negative reactions compared to other devices like mouse and keyboard (Buxton et al., 2007). In addition to these aspects, another problem that can interfere with Usability and/or UX is the lack of gesture standardization, which can require users to learn different gestures in different applications to perform the same task (Liang et al., 2011). Forlines et al. (2007) (Buxton et al., 2007) (Liang et al., 2011)

In SQ4, the results reveal that most technologies are being created for the study, which shows a lack of standardization among these technologies. This result corroborates the analysis shown in SQ3. Besides, even among the technologies based on existing ones, almost none consider the multi-touch context specificities.

With SQ5, we can affirm that the Likert scale is the most used collection method. It occurs 62 times, followed by log files with 17 occurrences, being the main methods to compose the quantitative data of SQ4. The main methods to gather qualitative data are composed of technologies that have observations, notes, and open-ended questions. Qualitative data evaluation has significant importance in improving systems quality, allowing to identify problems of Usability and UX in multi-touch systems.

Regarding SQ6, the characteristics of the evaluation technologies were verified and collected. Questionnaires created for the studies show their versatility, enabling any number and nature of questions to be made, allowing more autonomy for the authors in this process. However, we noticed the need and work of the authors to create an evaluation method for a single study.

About SQ7, results show that the most used analysis type is quantitative, being qualitative and both a small part of it. This result demonstrates that the author prefers to focus on numbers over subjective aspects in qualitative analysis. We believe that evaluation technologies that focus on quantitative and qualitative data can bring more diversity and points of view about the topics researched.

In SQ8, we observed that multi-touch systems perform several functions, from robot piloting to protein manipulation. This variety of functions is reflected in the presence of social areas, from accessibility to space exploration.

According to SQ9, we can observe that the primary goal

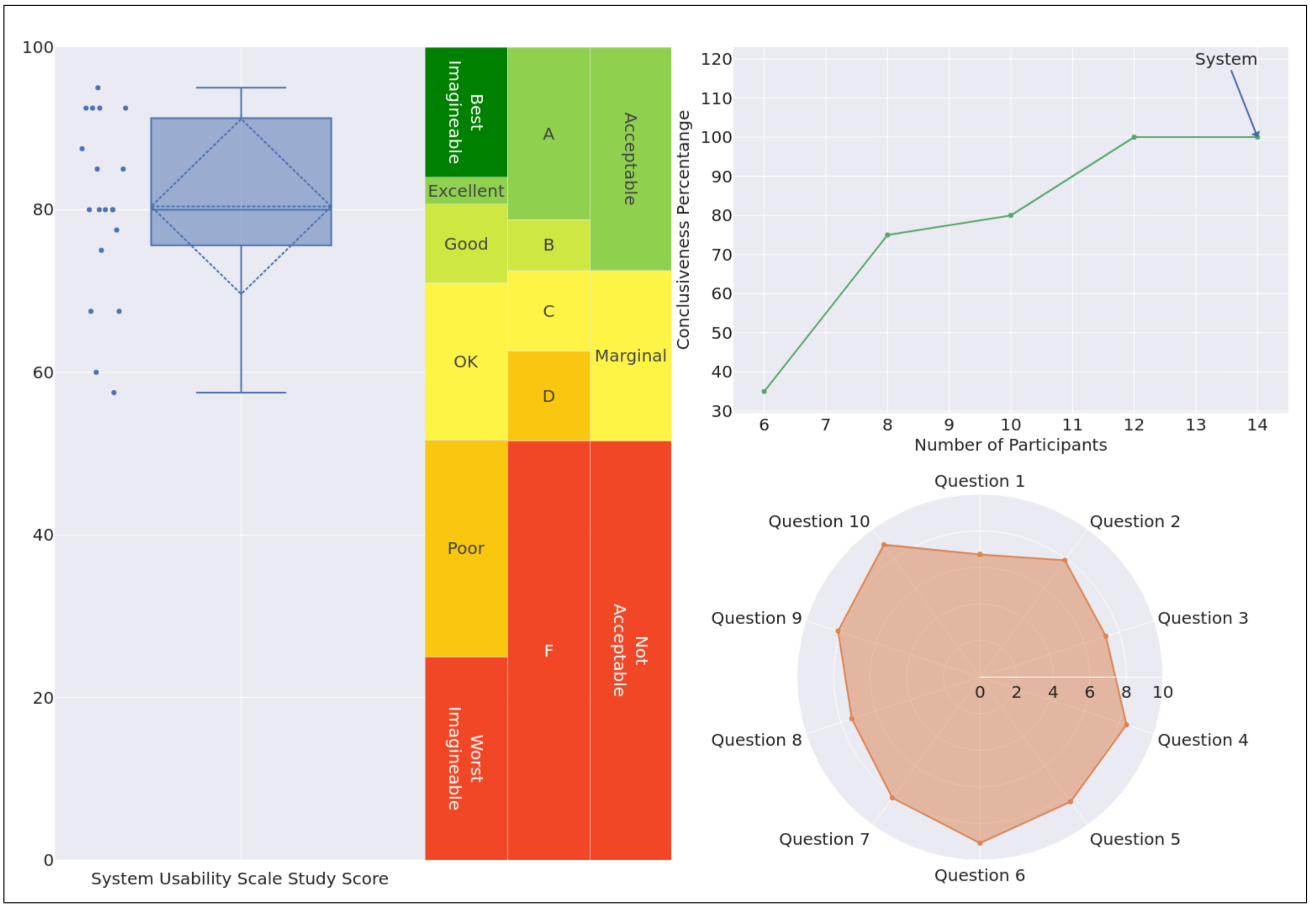

Figure 13. Three types of charts generated in one of SUS Analysis Toolkit presets.

of multi-touch systems is to facilitate the user's daily lives. The facilitation can be exemplified by expanding the use of multi-touch in several devices, such as smartphones, tablets, tabletops, and all sorts of screens. Our results also identify multi-touch systems in areas as mobility, health, entertainment, and other.

We observe that multi-touch systems are not restricted to general facilities, but they have a particular motivation in some cases. An example of motivation can be a group of people. Most of the multi-touch systems analyzed are not designed to help a specific group, but the results of SQ10 show that some have a group direction. In addition, we observed a link between multi-touch systems and accessibility since some of these systems focus on people with difficulty performing conventional interactions, such as children, blind and older adults.

About SQ11, almost all systems evaluated use 2 fingers to make the interaction. This result was expected since most of the multi-touch gestures known use 2 fingers. However, the use of 3 fingers also is common. Some studies show the use of 2 or more fingers in the experiment but do not make clear the exact number. Most of the interactions with more than 3 fingers were in systems that allow interaction with 10 fingers. Some interactions involved more than one person interacting at the same time. On these occasions, this paper analyzed only the interaction characteristics of a single participant.

With the SQ12, we verified the gestures used in the studies. We divided the gestures into six broad categories (drag, pinch, rotation, touch, touch and drag, free gesture) since each study can use a specific gesture that is not easily described. Besides, many studies give different names to the same gesture, mixing the functionality of the gesture with the gesture itself. From this, it was decided to group these gestures by the way they are materially performed. Drag is defined when the fingers touch the screen and go in any direction. Rotation occurs when two or more fingers perform a circular gesture, either around themselves or around an axis common to the fingers used. Touch and drag is defined by steady tapping of one finger while another finger is dragged across the screen. Touch is a simple touch on the screen. Pinch occurs when there is the movement of moving the fingers apart or closer, a gesture commonly used to perform the zoom action. Free gesture occurred only once, in a paper that allowed users to create any gesture for the interaction.

Regarding the SQs 13-15, all the technologies are used without being submitted to an empirical evaluation. The lack of empirical evaluation is confirmed by the high use of questionnaires created for only one study. Furthermore, according to our results, none of the evaluation technologies go through the empiricism process. This process seeks aspects such as verification of feasibility and validation, important steps to refine the technology and identify problems that can interfere with the quality of the evaluation (Mafra et al., 2006).

The SQs 16-18 attend the empirical evaluation of the multi-touch interfaces. A summary of the experiment description was collected, to allow a better understanding of the empirical process which the participants were submitted. This summary differ from paper to paper and can contain instructions given, tasks performed and environment descriptions. Also the amount of participants were collected. The study with the fewest participants had 2, while the largest had 326, and most of them do not surpasses 50 participants.

These results points to some directions that could be more explored in the future. Why so many technologies focus on extracting only quantitative data may be an important point in the field. The reason and motivation for the lack of Usability and UX assessment together can also be explored, and having the information that there is such a lack, the focus on filling it should be taken into account. It is also interesting to note that some concepts can be treated as UX by some authors and as Usability by others. The areas of HCI and SE would greatly benefit from work that sought to better establish a border between the concepts.

6.2 Benchmark discussion

We did the benchmark to provide an in-depth study of UX and usability questionnaires used in multi-touch context and identified in SMS. The need for complete systems to integrate all the necessary steps to conduct research with groups of people is remarkable. The usability and UX assessment technologies found and analyzed do not usually cover data distribution, collection, and analysis in the same tool, even for the most well-known questionnaires, such as the SUS.

Questionnaires like NASA-TLX do not cover the distribution and synthesis of results from more than one person in the same application. This lack of the same process with a beginning, middle, and end represented in a single technology may indicate that the usability and UX evaluation area is very niche, not making an effort to implement such a tool worthwhile. The lack of references and other analyses about these tools indicates that this lack prevented the perception of the non-existence of more complete tools.

7 Threats to Validity

It is shown in Ampatzoglou et al. (2019) the need and importance of reducing the risks to the validity of secondary studies. Therefore, some strategies were applied to reduce the impact of threats. We followed the well consolidated and solid protocol made by Kitchenham and Charters (2007). Since the search string is a determinant part of the results exhibited, field experts carefully reviewed their terms, and the string itself passed through a significant amount of tests to be refined.

Three researchers made part of this project, helping to reduce another threat, the authorial bias. This participation was in the process of protocol review, paper selection, extraction, and data analysis. The first researcher read all titles and abstracts and classified the papers of the first filter. The second and third researchers also read all titles and abstracts. A discussion between the three researchers was performed for each paper that did not agree on inclusion/exclusion. Next, the first researcher analyzed and extracted the papers' data in the second filter. The second and third researchers revised the extractions. Differences in inclusion or exclusion were discussed, and no paper was excluded without a plausible justification.

Threats that could not be controlled may have influenced our research results. For example, they are (i) insufficient information in papers to be extracted; (ii) non-standard definitions and treatments of terms; (iii) lack of detailing of evaluation technologies or multi-touch systems. However, we believe that the peer-review process reduced the impact of these threats.

8 Conclusions and Future Work

This paper accomplished an SMS to identify the Usability and UX evaluation technologies of multi-touch systems. Of 622 papers returned, 65 passed through the first and second filters and had their data extracted. Through this extraction, 29 different technologies were identified, and their characteristics were analyzed and classified by the SQs presented in Section 4.

Our results presented the authors' behavior in creating their evaluation technologies and metrics, most questionnaires, and log files. These technologies do not go under any process of empirical evaluation or validation. Besides, we noticed many evaluations that focus only on Usability. Our analysis also revealed that most technologies are generic, not considering the multi-touch specificities. Moreover, the majority do only quantitative analysis, not considering subjective thoughts that could contribute to the system quality. Our findings of multi-touch systems reveal their use to help people's daily lives, which do not dismiss their use in specific contexts such as accessibility and health.

The summary of our results can be divided into the following gaps: (i) few evaluation technologies for multi-touch systems that assess UX and Usability jointly; (ii) absence of empirical evaluation and validation process over the evaluation technologies; (iii) use of generic evaluation technologies do not consider the multi-touch specificities; (iv) most of the evaluation technologies have their focus only on quantitative analysis; (v) aspects of Usability and UX are not well defined in the context of multi-touch, with some of them being treated as both. We hope that these findings will help, inspire and collaborate with the scientific and industrial community in evaluating and creating new technologies to fill the gaps found.

The gaps identified by our SMS will serve as a basis to continue the work involving Usability and UX in the context of multi-touch systems quality. Also, the development of an evaluation technology that fills the gaps found can be conducted. We hope to contribute to the scientific community, industry, and society in this context. Besides, we expect that our SMS can serve as a basis for future SMSs.

The benchmark performed here demonstrated that there are few automated assessment technologies. After all, only 6 of the 14 questionnaires analyzed have online and/or automated versions. For future work, delivering automated tools for these questionnaires may be interesting. However, regarding new and existing automation, it is essential to highlight the need for a complete automated cycle of questionnaire distribution, data collection, and analysis. This step would be of fundamental importance to popularize the existing means of usability and UX evaluation and facilitate the use and application of these technologies.

Acknowledgements

This paper is a extended version of the paper written by Filho et al. (2022). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Ackad, C. J., Collins, A., and Kay, J. (2010). Switch: exploring the design of application and configuration switching at tabletops. In *ACM International Conference on Interactive Tabletops and Surfaces*, pages 95–104.

Adelson, J. L. and McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point likert-type scale. *Educational and Psychological measurement*, 70(5):796–807.

Altman, D. G. (1990). Practical statistics for medical research. pages 406–409. CRC press.

Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., and Chatzigeorgiou, A. (2019). Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and Software Technology*, 106:201 – 230.

Anand, G. and Kodali, R. (2008). Benchmarking the benchmarking models. *Benchmarking: An international journal*, 15(3):257–291.

Arnaud, A., Corrégé, J.-B., Clavel, C., Gouiffès, M., and Ammi, M. (2016). Exploration of virtual environments on tablet: comparison between tactile and tangible interaction techniques. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 357–361.

Basili, V. R. and Rombach, H. D. (1988). Towards a comprehensive framework for reuse: A reuse-enabling software evolution environment. In *NASA, Goddard Space Flight Center, Proceedings of the Thirteenth Annual Software Engineering Workshop*, number UMIACS-TR-88-92.

Berkman, M. I. and Karahoca, A. (2012). A direct touch table-top display as a multi-user information kiosk: Comparing the usability of a single display groupware either by a single user or people cooperating as a group. *Interacting with Computers*, 24(5):423–437.

Bertolo, D., Vivian, R., and Dinet, J. (2013). A set of interactions to rotate solids in 3d geometry context. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 625–630.

Blattgerste, J., Behrends, J., and Pfeiffer, T. (2022). A Web-Based Analysis Toolkit for the System Usability Scale.

Blažica, B., Vladušič, D., and Mladenić, D. (2013). Mti: A method for user identification for multitouch displays. *International Journal of Human-Computer Studies*, 71(6):691–702.

Brooke, J. (1996). Sus: A "quick and dirty" usability scale. In JORDAN, P. W. and *et al.*, editors, *Usability evaluation in industry*, pages 1–7, London. Taylor&Francis.

Buxton, B. et al. (2007). Multi-touch systems that i have known and loved. *Microsoft Research*, 56:1–11.

Campos, T. P. d., Damasceno, E. F., and Valentim, N. M. C. (2022). Proposal and evaluation of a collaborative is to support systematic reviews and mapping studies. In *XVIII Brazilian Symposium on Information Systems*, pages 1–8.

Cascales-Martínez, A., Martínez-Segura, M.-J., Pérez-López, D., and Contero, M. (2016). Using an augmented reality enhanced tabletop system to promote learning of mathematics: A case study with students with special educational needs. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(2):355–380.

Chen, Z., Lin, J., Zhang, G., Yang, W., Li, G., and Huang, G. (2016). A multi-touch interface for natural and seamless image composition. In *2016 11th International Conference on Computer Science & Education (ICCSE)*, pages 796–801. IEEE.

Chyung, S. Y., Roberts, K., Swanson, I., and Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement*, 56(10):15–23.

Ciocca, G., Olivo, P., and Schettini, R. (2012). Browsing museum image collections on a multi-touch table. *Information systems*, 37(2):169–182.

Colley, A., Väyrynen, J., and Häkkilä, J. (2015). In-car touch screen interaction: Comparing standard, finger-specific and multi-finger interaction. In *Proceedings of the 4th international symposium on pervasive displays*, pages 131–137.

Coram, J. L., Iverson, R., and Ackerman, A. (2013). Astrotouch: a multi-touch digital desktop for astrodynamics. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*, pages 11–14.

Damaraju, S., Seo, J. H., Hammond, T., and Kerne, A. (2013). Multi-tap sliders: advancing touch interaction for parameter adjustment. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 445–452.

de Souza Alcantara, T., Bastianelli, P., Ferreira, J., and Maurer, F. (2012). A multi-touch approach to control mri scans: A user-centered study report. In *2012 4th International Workshop on Software Engineering in Health Care (SEHC)*, pages 65–68. IEEE.

Derboven, J., De Roeck, D., and Verstraete, M. (2012). Semiotic analysis of multi-touch interface design: The mutable case study. *International Journal of Human-Computer Studies*, 70(10):714–728.

Dodd, C., Athauda, R., and Adam, M. (2017). Designing user interfaces for the elderly: a systematic literature review. pages 1–12. Association for Information Systems.

Ducasse, J., Macé, M., Oriola, B., and Jouffrais, C. (2018). Botmap: Non-visual panning and zooming with an actuated tabletop tangible interface. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(4):1–42.

Echtler, F., Nestler, S., Dippon, A., and Klinker, G. (2009). Supporting casual interactions between board games on public tabletop displays and mobile devices. *Personal and ubiquitous computing*, 13(8):609–617.

Fabroyir, H. (2019). Multitouch interface is not good for spatial navigation in virtual reality. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, pages 323–326. IEEE.

Fernandez, R. A. S., Sanchez-Lopez, J. L., Sampedro, C., Bavle, H., Molina, M., and Campoy, P. (2016). Natural user interfaces for human-drone multi-modal interaction. In *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1013–1022. IEEE.

Filho, G. E. K., Guerino, G. C., and Valentim, N. M. (2022). A systematic mapping study on usability and user experience evaluation of multi-touch systems. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, pages 1–12.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22(5):323–327.

Fleh, S. Q., Bayat, O., Al-Azawi, S., and Uçan, O. N. (2018). A systematic mapping study on touch classification. pages 1–9.

Forecast, M. D. (2022). Smartphone market size and growth. https://www.marketdataforecast.com/market-reports/smartphone-market.

Forlines, C., Wigdor, D., Shen, C., and Balakrishnan, R. (2007). Direct-touch vs. mouse input for tabletop displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 647–656.

Freitag, G., Tränkner, M., and Wacker, M. (2012). Enhanced feed-forward for a user aware multi-touch device. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pages 578–586.

Frisch, M., Langner, R., and Dachselt, R. (2011). Neat: a set of flexible tools and gestures for layout tasks on interactive displays. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, pages 1–10.

Fu, C.-W., Goh, W.-B., and Ng, J. A. (2010). Multi-touch techniques for exploring large-scale 3d astrophysical simulations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2213–2222.

Ghomi, E., Huot, S., Bau, O., Beaudouin-Lafon, M., and Mackay, W. E. (2013). Arpège: learning multitouch chord gestures vocabularies. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*, pages 209–218.

Giesler, A., Valkov, D., and Hinrichs, K. (2014). Void shadows: multi-touch interaction with stereoscopic objects on the tabletop. In *Proceedings of the 2nd ACM symposium on Spatial user interaction*, pages 104–112.

Glonek, G. and Pietruszka, M. (2012). Natural user interfaces (nui). pages 1–19. Politechnika Łódzka.

Guerino, G. C. and Valentim, N. M. C. (2020). Usability and user experience evaluation of natural user interfaces: a systematic mapping study. *IET Software*, 14(5):451–467.

Gürlük, H., Jauer, M.-L., and Uebbing-Rumke, M. (2014). Design and evaluation of a multi-touch interaction language for approach controllers. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*, pages 1–4.

Hachet, M., Bossavit, B., Cohé, A., and de la Rivière, J.-B. (2011). Toucheo: multitouch and stereo combined in a seamless workspace. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 587–592.

Hart, S. G. (1986). Nasa task load index (tlx). pages 1–26. Human Performance Research Group NASA Ames Research Center Moffett Field.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human–Computer Interaction*, 19(4):319–349.

Hassenzahl, M. and Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, 25(2):91–97.

Hsiao, H.-S., Chang, C.-S., Lin, C.-Y., Chang, C.-C., and Chen, J.-C. (2014). The influence of collaborative learning games within different devices on student's learning performance and behaviours. *Australasian Journal of Educational Technology*, 30(6).

Huerta, O., Sánchez, J. A., Fuentes, S., and Cervantes, O. (2011). Speak up your mind: Using speech to capture innovative ideas on interactive surfaces. In *Proceedings of the 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction*, pages 202–211.

ISO 9241-210 (2019). *Ergonomics of Human System Interaction - Part 210: Human-Centered Design for Interactive Systems*. International Organization for Standardization.

ISO 9241-9 (2000). *Ergonomics of Human System Interaction - Part 9: Requirements for non-keyboard input devices*. International Organization for Standardization.

ISO/IEC 25010 (2011). *Systems and Software Engineering - SquaRE - Software product Quality Requirements and Evaluation: System and Software Quality Models)*. International Organization for Standardization.

Jacucci, G., Morrison, A., Richard, G. T., Kleimola, J., Peltonen, P., Parisi, L., and Laitinen, T. (2010). Worlds of information: designing for engagement at a public multi-touch display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2267–2276.

Jetter, H.-C., Gerken, J., Zöllner, M., Reiterer, H., and Milic-Frayling, N. (2011). Materializing the query with facet-streams: a hybrid surface for collaborative search on tabletops. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3013–3022.

Johns, R. (2005). One size doesn't fit all: Selecting response scales for attitude items. *Journal of Elections, Public Opinion & Parties*, 15(2):237–264.

Kildal, J., Lucero, A., and Boberg, M. (2013). Twisting touch: combining deformation and touch as input within the same interaction cycle on handheld devices. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 237–246.

Kim, S. J., Ng, H., Winkler, S., Song, P., and Fu, C.-W. (2012). Brush-and-drag: A multi-touch interface for photo triaging. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 59–68.

Kim, Y.-H. and Lee, J.-H. (2015). Game interface enhancement under smartphone platform focused on touchscreen interaction. *Computers & Industrial Engineering*, 80:45–61.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Koutlemanis, P., Ntelidakis, A., Zabulis, X., Grammenos, D., and Adami, I. (2013). A steerable multitouch display for surface computing and its evaluation. *International Journal on Artificial Intelligence Tools*, 22(06):1360016.

Kulas, J. T. and Stachowski, A. A. (2013). Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on likert personality indicators. *Journal of Research in Personality*, 47(4):254–262.

Kulik, A., Kunert, A., Keil, M., and Froehlich, B. (2018). Rst 3d: A comprehensive gesture set for multitouch 3d navigation. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 363–370. IEEE.

Lammel, F. E., Manssour, I. H., and Silveira, M. S. (2016). Multi-touch interaction with information visualization techniques: An analysis using twitter data. In *International Conference on Human-Computer Interaction*, pages 211–222. Springer.

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*, pages 63–76. Springer.

Leftheriotis, I. (2013). User authentication in a multi-touch surface: a chord password system. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1725–1730.

Leftheriotis, I., Giannakos, M. N., Chorianopoulos, K., and Jaccheri, L. (2015). Investigating the potential of a two-finger chord button in multi-touch applications. In *Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces*, pages 337–342.

Liang, S., Cameron, J., and Baciu, G. (2011). Generic gesture kernel modeling and its application with virtual garment design. In *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*, pages 279–286.

Lissermann, R., Huber, J., Schmitz, M., Steimle, J., and Mühlhäuser, M. (2014). Permulin: mixed-focus collaboration on multi-view tabletops. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3191–3200.

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

Madni, T. M., Nayan, Y. B., Sulaiman, S., Abro, A., and Tahir, M. (2016). Usability evaluation of orientation techniques for medical image analysis using tabletop system. In *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, pages 477–482. IEEE.

Mafra, S. N., Barcelos, R. F., and Travassos, G. H. (2006). Aplicando uma metodologia baseada em evidência na definição de novas tecnologias de software. In *Proceedings of the 20th Brazilian Symposium on Software Engineering (SBES 2006)*, volume 1, pages 239–254.

Mafra, S. N. and Travassos, G. H. (2006). Estudos primários e secundários apoiando a busca por evidência em engenharia de software. *Relatório Técnico, RT-ES*, 687(06).

Martin-SanJose, J.-F., Juan, M.-C., Mollá, R., and Vivó, R. (2017). Advanced displays and natural user interfaces to support learning. *Interactive Learning Environments*, 25(1):17–34.

Matell, M. S. and Jacoby, J. (1971). Is there an optimal number of alternatives for likert scale items? study i: Reliability and validity. *Educational and psychological measurement*, 31(3):657–674.

Merrad, W., Héloir, A., Kolski, C., and Krüger, A. (2022). Rfid-based tangible and touch tabletop for dual reality in crisis management context. *Journal on Multimodal User Interfaces*, 16(1):31–53.

Micire, M., Desai, M., Drury, J. L., McCann, E., Norton, A., Tsui, K. M., and Yanco, H. A. (2011). Design and validation of two-handed multi-touch tabletop controllers for robot teleoperation. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 145–154.

Mossel, A., Venditti, B., and Kaufmann, H. (2013). 3dtouch and homer-s: intuitive manipulation techniques for one-handed handheld augmented reality. In *Proceedings of the Virtual Reality International Conference: Laval Virtual*, pages 1–10.

Muender, T., Gulani, S. A., Westendorf, L., Verish, C., Malaka, R., Shaer, O., and Cooper, S. (2019). Comparison of mouse and multi-touch for protein structure manipulation in a citizen science game interface. *Journal of Science Communication*, 18(1):A05.

Nacher, V. and Jaen, J. (2015). Evaluating the accuracy of pre-kindergarten children multi-touch interaction. In *IFIP Conference on Human-Computer Interaction*, pages 549–556. Springer.

Nacher, V., Jaen, J., Navarro, E., Catala, A., and González, P. (2015). Multi-touch gestures for pre-kindergarten children. *International journal of human-computer studies*, 73:37–51.

Nayebi, F., Desharnais, J.-M., and Abran, A. (2012). The state of the art of mobile application usability evaluation. In *2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE.

Olwal, A., Feiner, S., and Heyman, S. (2008). Rubbing and tapping for precise and rapid selection on touch-screen displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 295–304.

Pfeiffer, M., Kern, D., Schöning, J., Döring, T., Krüger, A., and Schmidt, A. (2010). A multi-touch enabled steering wheel: exploring the design space. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3355–3360.

Przybylski, A. K., Murayama, K., DeHaan, C. R., and Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in human behavior*, 29(4):1841–1848.

Radhakrishnan, S., Lin, Y., Zeid, I., and Kamarthi, S. (2013). Finger-based multitouch interface for performing 3d cad operations. *International Journal of Human-Computer Studies*, 71(3):261–275.

Rädle, R., Jetter, H.-C., Butscher, S., and Reiterer, H. (2013). The effect of egocentric body movements on users' navigation performance and spatial memory in zoomable user interfaces. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*, pages 23–32.

Renzi, A. B. and Freitas, S. (2014). Affordances and gestural interaction on multi-touch interface systems: building new mental models. In *International Conference of Design, User Experience, and Usability*, pages 615–623. Springer.

Rodriguez-Conde, I. and Campos, C. (2020). Towards customer-centric additive manufacturing: making human-centered 3d design tools through a handheld-based multi-touch user interface. *Sensors*, 20(15):4255.

Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68.

Ryan, R. M., Rigby, C. S., and Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, 30(4):344–360.

Santos, G., Rocha, A. R., Conte, T., Barcellos, M. P., and Prikladnicki, R. (2012). Strategic alignment between academy and industry: a virtuous cycle to promote innovation in technology. In *2012 26th Brazilian Symposium on Software Engineering*, pages 196–200. IEEE.

Schürmann, T., Binder, C., Janzarik, G., and Vogt, J. (2015). Movement transformation on multi-touch devices: Intuition or instructional preparation? *Applied Ergonomics*, 50:251–255.

Shneiderman, B. (1981). Direct manipulation: A step beyond programming languages. In *Proceedings of the Joint Conference on Easier and More Productive Use of Computer Systems.(Part-II): Human Interface and the User Interface-Volume 1981*, page 143.

Shull, F., Carver, J., and Travassos, G. H. (2001). An empirical methodology for introducing software processes. *ACM SIGSOFT Software Engineering Notes*, 26(5):288–296.

Telkenaroglu, C. and Capin, T. (2013). Dual-finger 3d interaction techniques for mobile devices. *Personal and ubiquitous computing*, 17(7):1551–1572.

Tseng, P.-H., Hung, S.-H., Chiang, P.-Y., Yao, C.-Y., and Chu, H.-K. (2018). Ez-manipulator: Designing a mobile, fast, and ambiguity-free 3d manipulation interface using smartphones. *Computational Visual Media*, 4(2):139–147.

Tuddenham, P., Kirk, D., and Izadi, S. (2010). Graspables revisited: multi-touch vs. tangible input for tabletop displays in acquisition and manipulation tasks. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2223–2232.

Tuveri, E., Iacolina, S. A., Sorrentino, F., Spano, L. D., and Scateni, R. (2013). Controlling a planetarium software with a kinect or in a multi-touch table: a comparison. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pages 1–4.

Uebbing-Rumke, M., Gürlük, H., Jauer, M.-L., Hagemann, K., and Udovic, A. (2014). Usability evaluation of multi-touch displays for tma controller working positions. *Proceedings of the 4th SESAR Innovation Days, Madrid, Spain*, pages 25–27.

Ullrich, D. and Diefenbach, S. (2010). From magical experience to effortlessness: an exploration of the components of intuitive interaction. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pages 801–804.

Väätäjä, H., Koponen, T., and Roto, V. (2009). Developing practical tools for user experience evaluation: a case from mobile news journalism. In *European Conference on Cognitive Ergonomics: Designing beyond the Product—Understanding Activity and User Experience in Ubiquitous Environments*, pages 1–8.

Vetere, F., O'Hara, K., Paay, J., Ploderer, B., Harper, R., and Sellen, A. (2014). Social nui: social perspectives in natural user interfaces. In *Proceedings of the 2014 companion publication on Designing interactive systems*, pages 215–218.

Vorderer, P., Wirth, W., Gouveia, F. R., Biocca, F., Saari, T., Jäncke, L., Böcking, S., Schramm, H., Gysbers, A., Hartmann, T., et al. (2004). Mec spatial presence questionnaire. volume 18, page 2015.

Wagner, J., Huot, S., and Mackay, W. (2012). Bitouch and bipad: designing bimanual interaction for hand-held tablets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2317–2326.

Wang, J. and Lindeman, R. (2012). Leaning-based travel interfaces revisited: frontal versus sidewise stances for flying in 3d virtual spaces. In *Proceedings of the 18th ACM symposium on Virtual reality software and technology*, pages 121–128.

Wang, J. and Lindeman, R. (2014). Coordinated 3d interaction in tablet-and hmd-based hybrid virtual environments. In *Proceedings of the 2nd ACM symposium on Spatial user interaction*, pages 70–79.

Wang, J. and Lindeman, R. (2015). Coordinated hybrid virtual environments: Seamless interaction contexts for effective virtual reality. *Computers & Graphics*, 48:71–83.

Watson, D. and Clark, L. A. (1994). The panas-x: Manual for the positive and negative affect schedule-expanded form. pages 1–28. University of Iowa.

Watson, D., Hancock, M., Mandryk, R. L., and Birk, M. (2013). Deconstructing the touch experience. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*, pages 199–208.

Weems, G. H. and Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and evaluation in counseling and development*, 34(3):166–176.

Wigdor, D. and Wixon, D. (2011). Brave nui world: designing natural user interfaces for touch and gesture. pages 1–156. Elsevier.

Wingert, B., Schöllhorn, I., and Bues, M. (2017). Prodesk: An interactive ubiquitous desktop surface. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, pages 366–371.

Witmer, B. G. and Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3):225–240.

Zaharias, P., Michael, D., and Chrysanthou, Y. (2013). Learning through multi-touch interfaces in museum exhibits: An empirical investigation. *Journal of Educational Technology & Society*, 16(3):374–384.

Zhang, C., Sahoo, D. R., Pearson, J., Robinson, S., Holton, M. D., Hopkins, P., and Jones, M. (2020). Active pinscreen: Exploring spatio-temporal tactile feedbackfor multi-finger interaction. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–11.