Evaluating Gesture User Interfaces: Quantitative Measures, Qualitative Scales, and Method*

Quentin Sellier[a,*], Arthur Sluÿters[b], Jean Vanderdonckt[b] and Ingrid Poncin[a]

[a]*Université catholique de Louvain, Louvain Research Institute in Management and Organizations (LouRIM), Chaussée de Binche 151, B-7000 Mons, Belgium*

[b]*Université catholique de Louvain, Louvain Research Institute in Management and Organizations (LouRIM), Place des Doyens 1, B-1348 Louvain-la-Neuve, Belgium*

ARTICLE INFO

Keywords:
Discoverability
Gesture interaction
Gesture user interface
Memorability
Multimedia browsing
Social Acceptability
Usability
User experience

ABSTRACT

Although many methods currently exist to evaluate a user interface, they have been mainly developed and applied for graphical and vocal user interfaces, leaving aside other modalities such as gesture interaction. As a consequence, the evaluation of the global quality of a gesture user interface most often resorts to these methods, which take little or no explicit and specific account of the gesture modality or which adapt existing methods in a way that jeopardizes their validity. To remedy this situation, this paper introduces, defines and justifies a method for evaluating explicitly and specifically a gesture user interface based on a conceptual model that consists of: (1) six quantitative measures covering user and system aspects, such as gesture thinking time and recognition rate; (2) three new qualitative scales, *i.e.,* discoverability, learnability, and social acceptability, combined with seven scales in a gesture evaluation scheme, which is formally based on four measures, *i.e.,* subscale score, subscale importance, scale mean score and scale mean importance; (3) a debriefing interview to cross-analyze the results of the quantitative measures and qualitative scales. This method also includes the use of multiple sessions to take into account short and long-term memory, as well as discoverability, and integrates tests specific to each gesture, as well as concrete use cases. For validation and illustration purposes, this method is applied to a gesture user interface serving as a case study. The results are then analyzed, starting with classic methods and gradually adding the particularities of our method to highlight its added value. Based on this experiment, we suggest some implications for evaluating gesture interfaces and for incorporating these scales into UEQ+, a modular user interface evaluation method.

1. Introduction

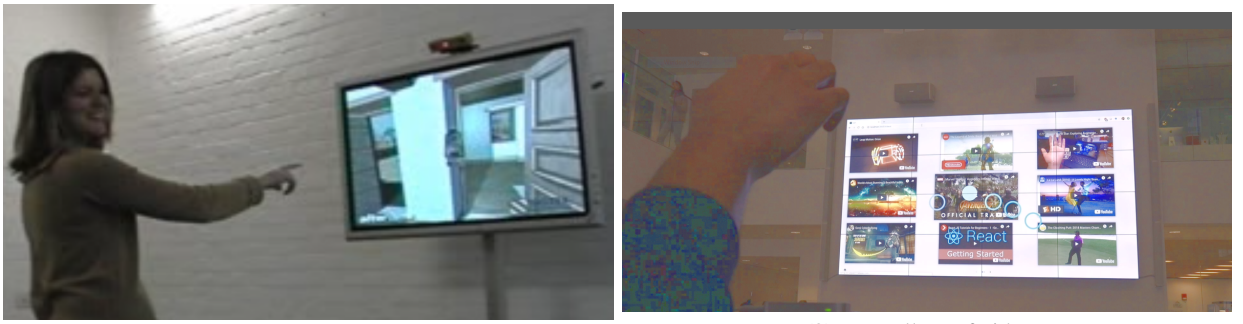
Many methods exist for evaluating the quality of a user interface (UI) in general (Dix, Finlay, Abowd and Beale, 2004; Vermeeren, Law, Roto, Obrist, Hoonhout and Väänänen-Vainio-Mattila, 2010), but not that many are specifically tailored to gesture UIs, a UI that brings into play the rich palette of all human movements for interacting with a system via sophisticated devices and sensors, such as a radar Siean, Pamparau, Sluÿters, Vatavu and Vanderdonckt (2023). For example, Fig. 1 shows a gesture UI to browse a 3D virtual reality scene and a gallery of videos. Given the particularities of gestural interaction compared to traditional interaction, some authors have already put forward the need for a more advanced evaluation method for gestural interaction (Wickeroth, Benölken and Lang, 2009), evaluating both qualitative and quantitative aspects (Farhadi-Niaki, Etemad and Arya, 2013), and capturing the multi-faced perspective of the richness of interaction (Xia, Glueck, Annett, Wang and Wigdor, 2022). Rohrer (2014)'s classification distinguishes between methods that produce qualitative measures that capture the quality attributes of the attitudinal UI, such as preference, and methods that produce quantitative measures that express the behavioral quality attributes of the UI, such as performance. This distinction is important since attitudinal measures represent the self-reported measures of

* The authors of this paper acknowledge the support of the MIT-Belgium MISTI Program under grant COUHES n°1902675706. Quentin Sellier is supported by the "Fonds Spéciaux de Recherche" under Grant reference ARH/LJB/01137236. Arthur Sluters is supported by a "mandat d'aspirant" of the "Fonds de la Recherche Scientifique - FNRS" under grants n°40001931 and n°40011629. Jean Vanderdonckt is supported by the EU EIC Pathfinder-Awareness Inside challenge "Symbiotik" project (1 Oct. 2022-31 Dec. 2026) under Grant n°101071147

*Corresponding author

quentin.sellier@uclouvain.be (Q. Sellier); arthur.sluyters@uclouvain.be (A. Sluÿters); jean.vanderdonckt@uclouvain.be (J. Vanderdonckt); ingrid.poncin@uclouvain.be (I. Poncin)

ORCID(s): 0000-0002-1379-0780 (Q. Sellier); 0000-0003-0804-0106 (A. Sluÿters); 0000-0003-3275-3333 (J. Vanderdonckt); 0000-0002-4225-0118 (I. Poncin)



(1) In a 3D Scene. (2) In a gallery of videos.

Figure 1: Gesture User Interfaces for multimedia browsing.

end users when interacting with the UI while behavioral measures report what end users actually do when interacting. End users may report a preference that is radically different from what they actually do: this expresses the eternal divergence between the end users' preference (what they say) and their performance (what they actually do).

Most UI evaluation methods (Vermeeren et al., 2010) target USABILITY, one of the main software quality factors (ISO, 2019), which is decomposed into three subfactors (ISO, 2018): EFFECTIVENESS refers to task performance by expressing how accurately and completely the user achieved the goals (*e.g.,* measured by the task completion rate and error rate), EFFICIENCY refers to the amount of effort required to achieve the level of effectiveness when achieving the goals (*e.g.,* measured by the task completion time and fatigue), SATISFACTION refers to how comfortable the user feels while using the system (*e.g.,* assessed by a questionnaire for subjective satisfaction). Therefore, there are no subjective and affective measures that are essential to human nature.

But currently the evaluation of User eXperience (UX) (Hassenzahl, 2008) goes beyond usability (van Beurden, Ijsselsteijn and de Kort, 2012) because it is influenced by both the UI and its context of use (Calvary, Coutaz, Thevenin, Limbourg, Bouillon and Vanderdonckt, 2003), which itself includes the user and the tasks, the platforms, devices, and the working environment. The ISO (2018) 9241 standard defines UX as a "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service". According to this definition, UX could include all user emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors, and accomplishments that occur before, during, and after use. The decisive benefits of gesture UI are not only revealed through usability, but much more through the UX, which encompasses a broader class of quality factors, such as enjoyment, stimulation, and identification (van Beurden et al., 2012).

Frequently, gestures are modeled in laboratory settings where usability testing should be carried out to evaluate the extent to which the gesture interface can be used by specified users in a specified context of use (Calvary et al., 2003) to achieve specified goals with effectiveness, efficiency, and satisfaction. However, the wide diversity of gesture UIs makes it difficult to decide which quality factors should be taken into account in a usability test (Xia et al., 2022). This observation also holds for methods that evaluate the UX of gesture UIs. A thorough evaluation of a gesture UI should go beyond simply assessing its usability, namely, by evaluating its UX. However, most evaluation methods focus on usability or, less frequently, user experience (Guerino and Valentim, 2020).

In principle, these evaluation methods apply to any interface, whatever its modality, since they have been designed and presented as such. However, it is clear that these methods have been applied, tested, and validated on corpora of graphical UI, even if this has been done on a wide variety of contexts of use. The same is true for their interpretation: the results of the evaluation are mainly mapped to graphical interfaces. Therefore, reusing classical evaluation methods, whether for usability or user experience, is not sufficient to account for all the richness of possibilities offered by a gesture interface and does not specifically account for the quality factors specific to this interface (Xia et al., 2022). Such an evaluation may be incomplete, partial, or even inadequate.

Therefore, we benefit from several possibilities: (1) to redefine an evaluation method within the strict framework of a gesture UI; (2) to extend an existing method to take into account new aspects not previously covered; (3) to create an entirely new method devoted to a gesture UI. This last option seems to be utopian, because of the complexity that such an activity would entail and because it would mean forgetting all the knowledge acquired in the field of evaluation. Thus, a combination of these options seems more realistic. This article contributes to the area of evaluation of gesture user interface by defining a method for explicitly and specifically evaluating a gesture user interface based on a conceptual model combining six quantitative measures and a set of qualitative scales, among them *i.e.,* discoverability, learnability,

and social acceptability. Then, this paper will apply the method based on this conceptual model to a case study. It will analyze these results in a detailed way and see how these measures account for the quality of a gesture UI by taking advantage of the complementarity and convergence of approaches allowing us to capture quantitative and qualitative measurements.

2. Literature Review

In the area of **questionnaire-based evaluation**, the System Usability Scale (SUS) (Brooke, 1996) became a classical usability questionnaire consisting of ten 5-point Likert (1932) scales with questions like Q_1="I think that I would like to use this system frequently". Lewis and Sauro (2009) showed that SUS actually has two factors: USABILITY covered by eight items and LEARNABILITY covered by items 4 and 10. Since they have reasonable reliability (Cronbach's α=.91 and .70, respectively) and they correlate with the overall SUS and with each other, they can be used separately. While such general-purpose questionnaires are in principle applicable to any UI, they can be applied to gesture UI too, but without highlighting particular aspects that are specifically relevant to the gesture modality. For example, Bhuiyan and Picking (2011) used a simplified version of the IBM Computer Satisfaction Usability Questionnaire (CSUQ) to evaluate the usability of Open Gesture, a hand-gesture UI that helps elderly people in everyday activities such as making phone calls, controlling their television, and performing mathematical calculations. In no case do the evaluation results highlight usability issues related to gesture interaction, since none of the questions deal with these aspects. Therefore, Wickeroth et al. (2009) extended SUS into a GESTURE USABILITY SCALE (GUS) by adding five items related to the gesture UI to evaluate the perceived reliability (G_1 and G_4), performance (G_3 and G_5), and compliance with the user's expectations (G_2):

G_1= "The system recognized the gestures reliably",
G_2= "The system followed the movements as expected",
G_3= "The reaction time of the system is satisfactory",
G_4= "The system aborted movements unexpectedly often", and
G_5= "The system reacted to my movements too slowly".

Although the GUS score was lower than the SUS score, a correlation existed between them. This method involves only self-reported items and does not compute any quality model. To overcome these shortcomings, Barclay, Wei, Lutteroth and Sheehan (2011) defined a quality model to evaluate a gesture UI as follows: accuracy, fatigue, and duration are objectively measured and naturalness is assessed using a 5-point Likert scale; their values are repeated for each gesture, normalized, averaged, and aggregated into a model where the weights of the factors are determined by participants. This results in a quantitative, yet simple, model for evaluating a gesture UI.

In order to compare a gesture UI with respect to a touch-based UI in a driving scenario, Graichen, Graichen and Krems (2019) combined questionnaire-based evaluation (in this case, a 12-point rating scale for trust, a driver distraction questionnaire, the NASA TLX questionnaire (Hart and Staveland, 1988) and the AttrakDiff questionnaire (Hassenzahl, 2008)) with quantitative data (in this case, glance points and duration). The questionnaires used are, again, widely applicable, therefore not targeting any particular modality such as gesture interaction. van Beurden et al. (2012) compared a gesture UI to a pointer-based UI, both in terms of their usability and UX through two experiments, one in a near-field context and one in a far-field context of use. Whereas pointer-based UIs scored higher in perceived performance and the mouse scored higher in pragmatic quality, ease of learning, and pragmatic quality, gesture UIs scored higher in terms of hedonic quality and fun, which demonstrates the need to evaluate a gesture UI beyond usability.

In the area of **heuristic evaluation**, Norman and Nielsen (2010), regretting that no valid usability guidelines exist yet for guidelines review, recommend the use of several heuristics, *i.e.,* visibility, feedback, consistency, discoverability, scalability, and reliability, but without going any further. Similarly, Wachs, Kölsch, Stern and Edan (2011) expressed eight heuristics specifically tailored for hand gestures, *i.e.,* price, responsiveness, user adaptability and feedback, learnability, accuracy, low mental load, intuitiveness, and comfort, each associated with a challenge. For example, accuracy poses challenges for the detection of gestures, its continuous tracking, and its real-time recognition, which are all complex problems to solve.

Later, Chuan, Sivaji and Ahmad (2014) proposed four general heuristics for evaluating a gesture UI: gesture learnability (to be evaluated before gesturing), gesture cognitive workload (to be evaluated during gesturing), gesture adaptability (to be evaluated after gesturing) and gesture ergonomics (to be evaluated when gesturing is prolonged over time). For example, gesture ergonomics should be assessed through comfort, fatigue, and physical ergonomics. Chuan,

Sivaji and Ahmad (2015) report that five evaluators discovered significantly more defects for a gesture UI with these heuristics than without them. While performing a heuristic evaluation based on these heuristics in addition to classical heuristics (Nielsen, 1994) is certainly desirable, no guidance is provided on how to effectively assess them, therefore basing the quality of this evaluation entirely on the expertise of the evaluators themselves and the knowledge they have of the domain. Moreover, the potential overlapping of these heuristics, taken together, could pose some problems of conflict (two different heuristics cover the same factor, but in a different way) or redundancy (two different heuristics cover the same factor in the same way).

In the area of **guidelines review**, a few usability guidelines are suggested to evaluate a gesture UI (Yee, 2009): "Achieve high effectiveness", "minimize learning among users and increase differentiation between gestures", and "design efficient gestures to increase user adoption". Again, these guidelines are hardly measurable per se and their sound evaluation is based on the seriousness of the evaluators and their experience. For example, how can we decide that two gestures are distinguishable enough if we cannot rely on a precise measure or a guideline that is empirically valid, with a strength of evidence and an importance level?

In the area of **user testing**, Farhadi-Niaki et al. (2013) compared 3 finger *vs.* 3 arm gestures for simple *vs.* complex desktop tasks using qualitative measures, such as easiness and fatigue, and quantitative measures, such as task completion time, to investigate potential correlations. While this combination enables a more thorough evaluation, it remains isolated and not put into perspective with other usability aspects. Fonseca Brandao, Casseb, Almeida, Assis, Camargo, Min and Castellano (2019) evaluated a gesture UI for a tele-rehabilitation system by relying on a functional magnetic resonance imaging (fMRI) protocol. By monitoring correlation maps between a region of interest in the primary motor cortex and brain voxels, they evaluated how brain connectivity changes over time before and after gesturing. While this evaluation method is quantitative, it is sophisticated, requires a specific setup, and is tied to brain activities, thus making it challenging to transpose to any gesture UI outside the area of tele-rehabilitation and to generalize. Neca and Duarte (2011) measured the average reflection time to manipulate images: in the "table of images" scenario, the average reflection time was larger in the gesture UI condition than in the gesture+vocal UI condition, whereas in the "wall of images" scenario, times were comparable.

3. Conceptual Model

This section defines our conceptual model for evaluating a gesture UI. First, we give the general principles of the protocol, dividing the experiment into two sessions evaluating short-term memory and long-term experience. We then detail the quantitative and objective measures of the interaction, the qualitative scales to use, the questionnaire, and the interview guidelines.

3.1. Protocol

The protocol divides the experiment into two sessions to properly evaluate the discoverability, as well as the short- and long-term memorability of the gesture set (Vogiatzidakis and Koutsabasis, 2022). This division also makes it possible to measure the skill acquisition and evaluate the learning process of users (Cockburn, Gutwin, Scarr and Malacria, 2014).

3.1.1. Session 1 (Discoverability and Short-term Memorability)

The objective of the first session is to introduce participants to the tested interface and to evaluate the user experience, discoverability, as well as short-term memorability of its gesture set (Sluÿters, Sellier, Vanderdonckt, Parthiban and Maes, 2022). It should include the following phases:

1. *Consent form and any demographic information.* The participants are introduced to the procedure and informed that they can choose to leave the experiment at any time. They are then invited to sign a consent form and fill in a demographic survey.
2. *Discovery phase.* Participants interact freely with the interface. They should receive the least instructions and information possible, just enough to know that they should use gesture interaction using specific sensors. This phase served to assess the discoverability of the gesture UI.
3. *Learning phase.* This phase serves to teach gestures to the participants and to let them familiarize themselves with the system. We recommend to use videos to ensure that all participants receive the same instructions.
4. *Resting phase.* We advise to give participants some time to rest in a separate room between these phases.

5. *Testing phase 1.* This phase aims to evaluate the short-term memorability of the gesture set by asking participants to perform specific tasks without external help. We recommend mixing task types with "atomic" tasks, requiring only a single gesture, and "compound" tasks, requiring a combination of gestures. The first category aims to focus on the ergonomics of each gesture, while the second simulates use cases closer to a real experience.
6. *Interview and questionnaires.* After completing the testing phase, participants should be interviewed to collect their feedback about the interface. Then, they should be asked to complete a questionnaire covering the selected scales for the interface. The interview questions can be quick but are mainly intended to let the participant express themselves on the interface, potentially expanding on their experience and letting them bring up points not covered by the questionnaires.

3.1.2. Session 2 (Long-term Memorability)

Session 2 is much shorter than the first session and should consist of the following phases:

1. *Introduction.* Participants are reminded of the procedure.
2. *Testing phase 2.* This second testing phase serves to evaluate the long-term memorability of the gesture set. We follow the same procedure as in the testing phase of the first session.
3. *Questionnaire.* At the end of the session, participants are asked to fill out the same scales as in the first session.

3.2. Measures

The conceptual model for evaluating a gesture user interface consists of quantitative and qualitative variables. The first part of the model is calculated objectively during the experiment and consists of the following dependent variables:

- GESTURE TASK SUCCESS RATE: a real variable between 0 and 100 that measures the percentage between the number of successful gesture tasks (*i.e.,* the gesture is correctly remembered, produced, and recognized) and the total number of executed gesture tasks.
- NUMBER OF GESTURE TRIALS: a positive integer that measures the total number of gestures executed for a given task, whether successful, incorrectly recalled, incorrectly executed, or incorrectly recognized by the gesture UI.
- GESTURE TASK COMPLETION TIME: a positive real variable that measures in seconds the time that a participant needs to complete a gesture task, which is the time elapsed between a stimulus and the completion of a gesture task. Note that a gesture task can involve more than one single gesture, *e.g.,* when repeated.
- GESTURE TASK AVERAGE COMPLETION TIME: a positive real variable that averages completion times for a set of participants for a given gesture task.
- GESTURE AGREEMENT RATE: a real variable between 0 and 100 that measures the agreement among a set of pairs of participants to elicit a particular gesture for a gesture task (Vatavu and Wobbrock, 2015).
- GESTURE RECOGNITION RATE: a real variable between 0 and 100 that measures the ratio between the number of correctly recognized gestures and the number of gesture trials. Note that a gesture could be properly remembered by a participant, but not accurately recognized by the gesture UI or could be inadequately remembered but accurately recognized.

The second part of the model consists of a GESTURE EVALUATION SCHEME, which is composed of a set of qualitative scales $S_i, \forall i \in \{1, \ldots, n\}$ decomposed in m_i subscales $SC_{ij}, \forall j \in \{1, \ldots, m_i\}$ or items to be evaluated (*e.g.,* the scale ATTRACTIVENESS of UEQ+ is decomposed in four subscales: annoying *vs.* enjoyable, bad *vs.* good, unpleasant *vs.* pleasant, and unfriendly *vs.* friendly), each subscale SC_{ij} being a differential scale with 7 points between items of each pair (*e.g.,* annoying o o o o o o o enjoyable). We measure each subscale SC_{ij} using a n point Likert scale with response categories, such as 1="Strongly disagree" to 7="Strongly agree" (=7). For each S_i, we define:

1. SUBSCALE SCORE (SS): an ordinal integer variable that measures the score for each subscale SC_{ij} of a scale S_i, ranging from a lower to a higher bound, such as 1="Strongly disagree" to 7="Strongly agree" for 7 points.
2. SUBSCALE IMPORTANCE (SI): an ordinal integer variable that measures the weight for each subscale SC_{ij} of a scale S_i, ranging from a lower bound to a higher bound, such as 1="Least important" to 7="Most important".
3. SCALE MEAN SCORE (SMS): a real variable that measures the average score obtained on all subscales SC_{ij} of a scale SC_i for all participants, ranging from -3="Mostly negative" to +3="Mostly positive" computed as follows: $\mathrm{SMS}(SC_i) = \frac{\sum_{j=1}^{m} \mathrm{SS}(SC_{ij})}{n} - \mathrm{Mdn}, \forall i \in \{1, \ldots, n\}$, where Mdn=4 is the scale median.

4. SCALE MEAN IMPORTANCE (SMI): a real variable that measures the average weight of importance of all subscales SC_{ij} of a scale SC_i for all participants, ranging from -3="Mostly negative" to +3="Mostly positive" computed as follows: $\text{SMI}(SC_i) = \frac{\sum_{j=1}^{m} \text{SS}(SI_{ij})}{n} - \text{Mdn}, \forall i \in \{1, \dots, n\}$, where Mdn=4 is the median for a 7-point scale.

Regarding the interviews, all sessions should be conducted by the same experimenter to ensure consistency (Riedel, Weeks and Beatson, 2018). Once the data collection phase is over, we can use thematic analysis to analyze and interpret the data corpus. Thematic analysis is a method that identifies, analyses, and reports patterns (themes) within data to obtain a rich and detailed description of it (Boyatzis, 1998; Braun and Clarke, 2006). The analysis is both deductive and inductive (Valos, Maplestone, Polonsky and Ewing, 2017) by adopting a "theory-driven" approach, with themes and codes identified in the prior literature and a "data-driven approach" with other codes emerging from the data (Braun and Clarke, 2006).

3.3. Scales

Xia et al. (2022) identified a set of 13 factors that are considered relevant and crucial for designing any gesture vocabulary. These factors are classified in four categories: situational (context, modality, social acceptability), cognitive (discoverability, intuitiveness, learnability, transferability), physical (complexity, efficiency, ergonomics, occlusion), and system (feedback, recognition). As it seems impossible and unrealistic to take all these thirteen criteria into account, it is the designer's responsibility to identify which criteria are important to take into account, the extent to which they should be prioritized and how they should be evaluated. Given the current state of research and literature, there is no single method that can meet this ongoing challenge. However, there are many studies reporting on experiences with one or other of these factors. But these studies usually focus on one criterion in isolation, which makes it difficult, if not impossible, to separate the considerations: studying one factor in isolation does not relate it to the other criteria. It is therefore difficult to differentiate between them and to study them holistically. It is not impossible that the study of one criterion, however rigorous it may be, also covers aspects linked to other criteria. This is often the case with questionnaires where we discover, after the fact, that their questions cover related aspects, that the answers given to one or other question also concern aspects related to another question. As it is virtually impossible to take all the criteria into account simultaneously, and as taking individual criteria into account, one after the other, is equally illusory (it is an elusive problem), we are moving towards a modular and incremental method. This doesn't mean that the criteria cannot be clearly separated, but it does mean that they can at least be taken into account in an integrated way.

In our current approach, we advise to select an empirically validated questionnaires such as UEQ+, a flexible and modular method to evaluate the UX of any UI (Schrepp and Thomaschewski, 2019). Based on these definitions, the usability could be evaluated through its three subfactors as follows: *Effectiveness* is covered by the GESTURE TASK SUCCESS RATE and GESTURE RECOGNITION RATE variables; *Efficiency* is covered by the NUMBER OF GESTURE TRIALS, GESTURE (AVERAGE) THINKING TIME, and GESTURE MEMORABILITY RATE; *Subjective satisfaction* is covered by the SOCIAL ACCEPTABILITY scale and by the debriefing interview. Any such scale could, of course, supplement any of the three subfactors. UX could be covered by any Gesture Evaluation Scheme and its related measures, such as the factors proposed by UEQ+ or the Pragmatic Quality (PQ), Hedonic Quality Stimulation (HQS), and Hedonic Quality Identification (HQI), as defined by Hassenzahl (2008). Regarding the specific scales of the UEQ+ questionnaire, we currently selected all the 7 scales that could be used in the context of gestural interaction (*i.e.,* ATTRACTIVENESS, EFFICIENCY, PERSPICUITY, DEPENDABILITY, STIMULATION, USEFULNESS, and INTUITIVE USE).

To include dimensions not present in UEQ+ but particularly important in the case of gestural interaction (Xia et al., 2022), we also introduce three qualitative scales. By doing so, our objective is not to strictly validate them but rather to offer a global view on these dimensions, also important in the case of gestural interaction, and potentially to open the conversation in debriefing interview. In the long term, it could be desirable to validate these scales, or even integrate them directly into UEQ+ in order to further standardize our method. We therefore introduce the following three qualitative scales :

- DISCOVERABILITY: refers to the ability to make the gestures straightforward for the end user to locate, observe, and discover how they work (Wobbrock, Morris and Wilson, 2009). It is an important aspect, as it enables one to quickly navigate in a UI without going through complete training (Mackamul, 2022). Nacenta, Kamber, Qiang and Kristensson (2013) argue that "[...] for users to leverage gestures in the first place, they have to discover and learn how to apply them effectively". Since gestures are by nature performed by a human movement

captured outside the UI but in the physical environment, the UI should reflect in some way that these gestures exist to execute the commands they are looking for. The effect of the action associated with a gesture is, of course, minimal feedback. Different mechanisms, such as those based on feed-up and feedforward, could let the end user discover the existence of these gestures and how to operate them. It is also called *guessability* or *approachability* (Xia et al., 2022). This scale is decomposed into three subscales:

1. D_1 ="I could easily guess how to use the gesture interface".
2. D_2 ="It was not complicated to find out how to control the gesture interface".
3. D_3 ="The gestures to control the gesture interface were easy to discover".

- LEARNABILITY: refers to the ease with which a new user can learn the interface and achieve optimal performance gestures (Dix et al., 2004). This factor is one of the most popular in the literature (Xia et al., 2022). It is related to intuitiveness as it is easier to learn how to use a user interface based on intuitive gestures. The ISO (2019) 25010 standard on software quality defines learnability as the "degree to which a product or system can be used by specified users to achieve specified goals of learning to use the product or system with effectiveness, efficiency, freedom from risk, and satisfaction in a specified context of use". There are different methods for assessing learnability, including the use of think-aloud or question-suggestion protocols (Grossman, Fitzmaurice and Attar, 2009). With the aim of standardizing our method, we favor the use of a qualitative scale. This scale, combined with debriefing interviews, also helps uncover possible learnability issues that would not have been found otherwise (Grossman et al., 2009). The specific scale that we use is decomposed into three subscales:

  1. L_1 ="It was easy to learn how the gesture interface works".
  2. L_2 ="The mastery of the gesture interface was not difficult to acquire".
  3. L_3 ="I felt a progression in my ability to control the gesture interface".

  This learnability scale is specifically adapted to disruptive modes of interaction and, in particular, gestural interaction. The third subscale refers to a progression and is therefore no longer relevant when the individual has perfectly mastered the method of gestural interaction.

- SOCIAL ACCEPTABILITY: refers to the UI ability to provide end users with gestures that are acceptable to be issued in a particular context of use. This aspect is essential to ensure the use of the interface in a public environment, but also to facilitate the discovery of gestures by new users who observe those who interact (Walter, Bailly and Müller, 2013). It depends on the social (Rico and Brewster, 2009) and cultural factors (Archer, 1997) that affect the experience when a gesture is performed. It can be influenced by the place where the interaction occurs, the potential audience, the age of the user, the gestures produced, etc. (Montero, Alexander, Marshall and Subramanian, 2010; Williamson, 2012). It "is determined when motivations to use technology compete with restrictions of social settings" (Rico and Brewster, 2009), therefore influenced by the context of use, *i.e.,* user and tasks, platforms, and environments (Calvary et al., 2003). This scale is decomposed into three subscales covering comfort (global well-being in a particular environment) and embarrassment (directly caused by one's actions):

  1. SA_1 ="I would feel comfortable interacting with the gesture interface in the presence of strangers".
  2. SA_2 ="I would have no problem using this gesture interface in the middle of a store".
  3. SA_3 ="I would not feel embarrassed to use this gesture interface in the presence of others".

4. Case Study: Evaluating a Gesture User Interface

An experiment was conducted following our procedure to evaluate the usability of LUI, as well as the short- and long-term memorability of its gesture set. This section details the design of this case study.

4.1. Interface

The LARGE USER INTERFACE (LUI) is a mid-air gesture-based interactive application that enables any general audience to manipulate multimedia objects (*i.e.,* photos, videos, documents and maps) on any display (*e.g.,* a video wall screen, a TV screen, a tabletop, or a video projector). It implements a wide set of multimedia actions, including zooming in/out a picture, rotating a document, and playing a video (Sluÿters et al., 2022). LUI and its gesture set were designed

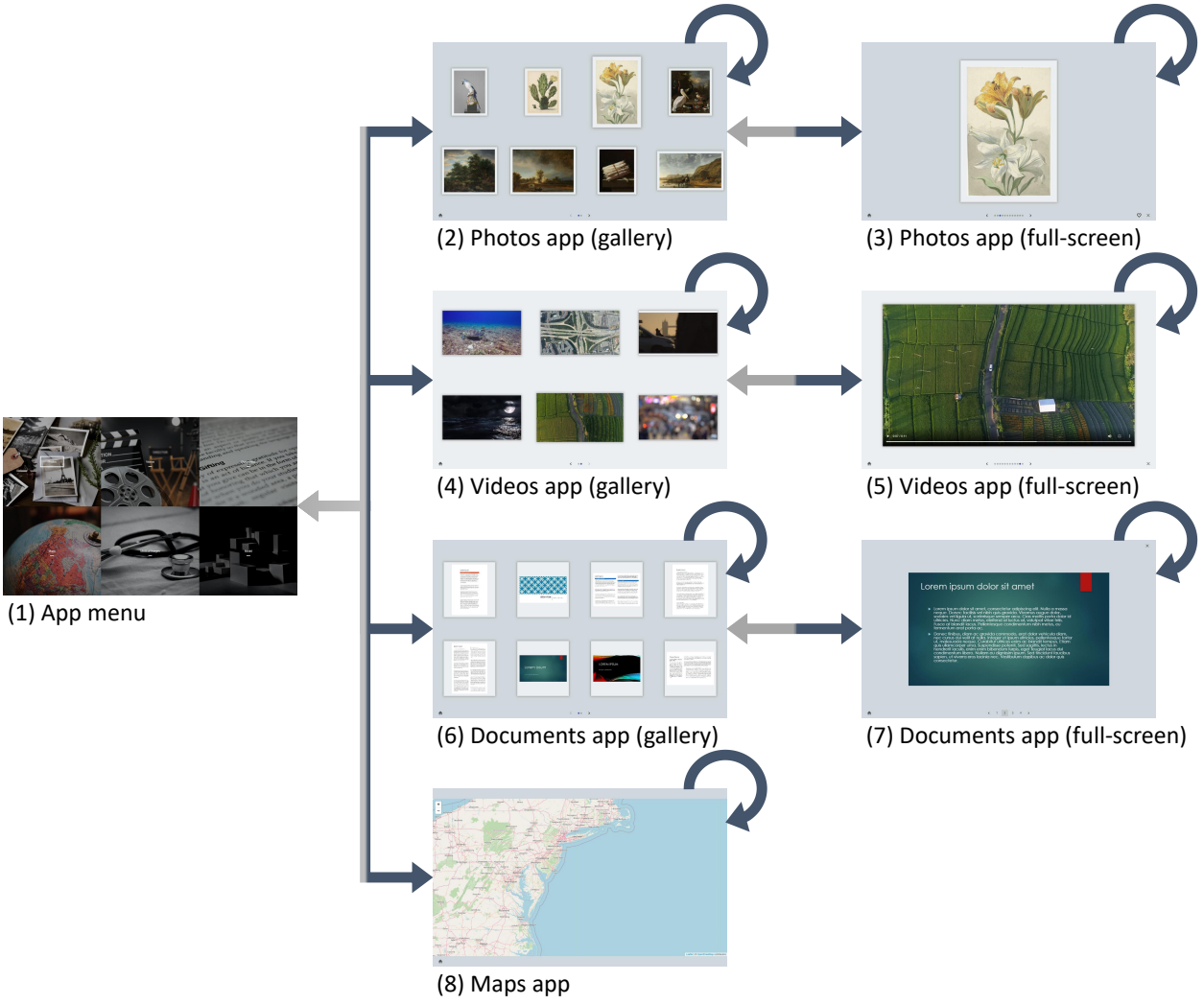
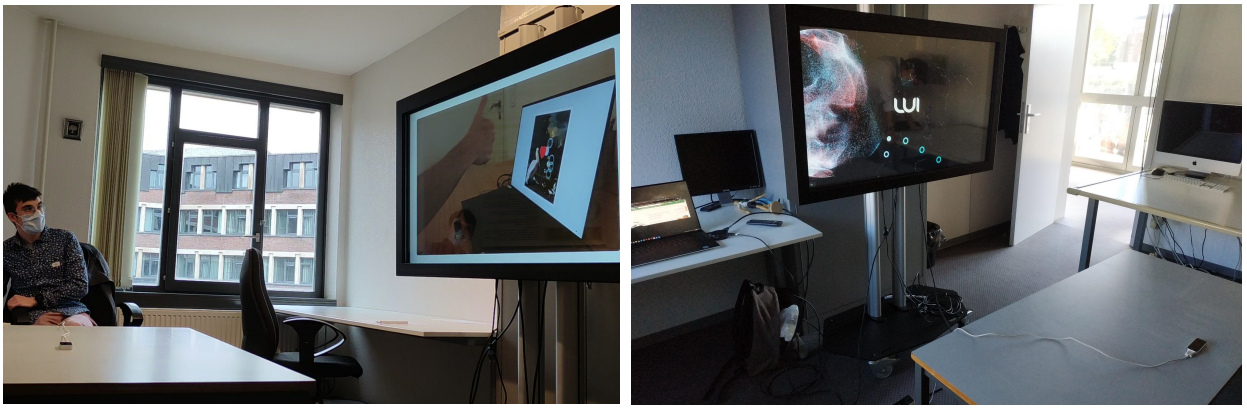


Figure 2: Screenshots of the last version of the LUI application. Arrows indicate the chaining between the pages of the UI.

to be highly learnable and intuitive to use, as well as to impose as few constraints as possible on the user. As such, the system relies on a Leap Motion Controller (LMC), an affordable and non-intrusive off-the-shelf device for capturing users' motion (Bachmann, Weichert and Rinkenauer, 2018). Thanks to this device, LUI automatically extracts raw data from the continuous stream of motion data without additional input from the user. LUI consists of four applications, *i.e.,* Photos, Videos, Documents, and Maps, which can be accessed through an application menu (Fig. 2-1) in which the user's fingers are projected as circles (Fig. 2-9) (Parthiban, Maes, Sellier, Sluÿters and Vanderdonckt, 2022).

4.2. Participants

Seventeen participants (8 female, 9 male, and 0 identified as gender variant/non-conforming) aged between 22 and 65 years old (M=37.5 years, SD=15.3 years) were randomly selected from a list of volunteers using stratified sampling to balance between gender, background, and age. All 17 participants used smartphones multiple times every day, and 16 participants (16/17=94.1%) used computers weekly, with 15 of them (15/17=88.2%) using computers multiple times per day. Two participants (2/17=11.8%) used gesture interaction multiple times a week. Participants' occupations included student, researcher, employee, computer scientist, physiotherapist, professor, and accountant. All participants were right-handed.



(1) Tutorial video playing on the screen. (2) Physical configuration.

Figure 3: The setup of the experiment.

4.3. Apparatus

Participants were seated at a desk and faced a large TV screen situated approximately two meters away, which displayed the LUI application and tutorial videos. An LMC was placed on the desk to capture their gestures. A laptop computer served to run the LUI application and control the experiment using custom software (*e.g.,* to play a tutorial video - see Fig. 3–1). In addition, a camera recorded both the screen and the participant's hands, and logs of the LUI application were saved on the system.

4.4. Protocol

As developed in Section 3.1, the experiment was divided into two sessions named $S1$ and $S2$ to properly evaluate the discoverability, as well as the short- and long-term memorability of the gesture set. The interval between two sessions was 7 days on average (SD=2.2 days).

4.4.1. Session 1 (Discoverability and Short-term Memorability)

The objective of the first session was to introduce participants to LUI and to evaluate the user experience, discoverability, as well as short-term memorability of its gesture set (Sluÿters et al., 2022).

1. *Consent form and demographic information.* The participants were introduced to the procedure and informed that they could choose to leave the experiment at any time. They were then invited to sign a GDPR-compliant consent form and fill in a demographic survey with information such as their age, gender, level of education, occupation, and previous experience with technologies such as smartphones, computers, and gesture interaction.
2. *Discovery phase.* Participants interacted freely with the LUI application for three minutes, starting from the App menu. They did not receive any specific instructions or information about the interface. The only explanation given to them is that it was controllable using the movements of their dominant hand above the sensor, without touching any surface.
3. *Learning phase.* The participants were given 17 tasks in a random order (Table 1). Each task was read aloud by the experimenter, who then played a short tutorial video of the associated gesture (Fig. 3–1). Once the tutorial finished playing, the experimenter loaded the associated page of the LUI application and prompted the participants to perform the gesture to complete the task (Fig. 3–2). Participants could re-watch the tutorial videos as many times as necessary. If a participant accidentally left the current page (*e.g.,* by performing the wrong gesture), the experimenter informed them and reloaded the correct page.
4. *Resting phase.* Participants were given five minutes to rest in a separate room.
5. *Testing phase 1.* As detailed in Section 3.1, the participants first completed 17 atomic tasks (Table 1) in a random order, following a similar procedure to the learning phase: the experimenter first read the task aloud, loaded the associated page of the LUI application, and prompted the participants to perform the gesture. The target content was randomly selected for each task (*e.g.,* for the "zoom in" task, the target could be a picture, a document, or a map), and the tasks were grouped by target content. The participants then performed three compound tasks (Table 2). For each compound task, the experimenter read the task aloud, loaded the application menu,

Table 1
The 17 atomic tasks, potential target content, associated gestures, and GESTURE AGREEMENT RATE *AR(r)* (Sluÿters et al., 2022)

Id	Task	Target(s)	Gestures	AR(r)
A1	Previous	Photo/video/document/page	Flick right	0.498
A2	Next	Photo/video/document/page	Flick left	0.498
A3	Enable full screen	Photo/video/document/app	Tap	0.273
A4	Disable full screen	Photo/video/document/app	Flick up	0.146
A5	Zoom in	Photo/document/map	Pinch out	0.32
A6	Zoom out	Photo/document/map	Pinch in	0.3
A7	Volume up	Video	Swipe up	0.478
A8	Volume down	Video	Swipe down	0.316
A9	Rotate clockwise	Photo/document	Rotate a knob clockwise	0.83
A10	Rotate anti-clockwise	Photo/document	Rotate a knob anti-clockwise	0.83
A11	Play	Video	Tap	0.19
A12	Pause	Video	Tap	0.166
A13	Like	Photo	Thumbs up	0.751
A14	Dislike	Photo	Thumbs down	0.751
A15	Fast-forward	Video	Swipe left	0.087
A16	Rewind	Video	Swipe right	0.095
A17	Pan	Map	Move hand while pointing index	0.433

Table 2
The three compound tasks and the minimum set of actions required to complete a task.

Id	Task	Instructions	Required gestures
C1	Maps	In the Maps app, move and enlarge the map so that the two blue markers are situated at the left and right sides of the screen.	Enable full screen, zoom in, pan
C2	Photos	In the Photos app, manipulate the windmill painting so that the author's name is the right way up and large enough to be easily readable.	Enable full-screen, next, zoom in, rotate clockwise and/or anticlockwise
C3	Videos	In the Videos app, find the words pronounced halfway through the "sea turtles" video. Fast forward through the video to get to the right part.	Enable full screen, next, play, fast-forward

and prompted the participants to perform the task. Participants could request the experimenter to repeat the instructions or to go back to the application menu.

6. *Interview and questionnaires.* After completing the testing phase, participants were interviewed to collect their feedback about the LUI application. The interview consisted of three open-ended questions: (1) "What is the first positive thing that comes to mind regarding your experience with the LUI application?", (2) "What is the first negative thing that comes to mind regarding your experience with the LUI application?", and (3) "Do you have any other comments?". Then, they were asked to complete a questionnaire consisting of seven evaluation scales selected from the UEQ+ method and our three new scales that target DISCOVERABILITY, LEARNABILITY, and SOCIAL ACCEPTABILITY of the LUI application.

4.4.2. Session 2 (Long-term Memorability)

Session 2 was much shorter than the first session and consisted of only three phases:

1. *Introduction.* Participants were welcomed and reminded of the procedure.
2. *Testing phase 2.* We followed the same procedure as in the testing phase of the first session.
3. *Questionnaire.* At the end of the session, participants were asked to fill out the same scales selected from UEQ+ (Schrepp and Thomaschewski, 2019) and the LEARNABILITY and SOCIAL ACCEPTABILITY scales as they did in the first session, but not the DISCOVERABILITY, which is no longer considered relevant.

Table 3
GESTURE TASK SUCCESS RATE for the 17 atomic and 3 compound tasks.

Id	Task	Learning phase	Testing phase 1	Testing phase 2	Average
A1	Previous	100%	100%	100%	100%
A2	Next	100%	94%	94%	96%
A3	Enable full screen	88%	82%	100%	90%
A4	Disable full screen	76%	88%	100%	88%
A5	Zoom in	100%	88%	94%	94%
A6	Zoom out	94%	94%	82%	90%
A7	Volume up	94%	88%	82%	88%
A8	Volume down	82%	100%	94%	92%
A9	Rotate clockwise	71%	71%	71%	71%
A10	Rotate anti-clockwise	82%	82%	71%	78%
A11	Play	82%	88%	100%	90%
A12	Pause	94%	88%	82%	88%
A13	Like	100%	100%	100%	100%
A14	Dislike	100%	100%	94%	98%
A15	Fast-forward	88%	88%	88%	88%
A16	Rewind	82%	82%	71%	78%
A17	Pan	88%	88%	88%	88%
C1	Photo	–	100%	94%	97%
C2	Video	–	82%	94%	88%
C3	Map	–	76%	65%	71%
	Average	89%	89%	88%	

5. Results and Discussion

In this section, we develop and discuss the main results of the evaluation carried out. We start with the quantitative measures, following a more traditional approach, to gradually add the qualitative aspects of our method and demonstrate the added value of including and combining them.

5.1. Quantitative Measures

This first part focuses on quantitative measurements, following a more traditional approach to interface evaluation. We are particularly interested in the success rate of the tasks, the number of trials, the time required to execute the tasks, and the recognition rate.

5.1.1. Gesture Task Success Rate

Table 3 shows the GESTURE TASK SUCCESS RATE for the atomic and compound tasks for the learning and testing phases. On average, participants succeeded in 89% of their gestural tasks, with remarkable consistency, starting with 89% during the learning phase, maintaining the rate for the first testing phase, and slightly decreasing to 88% during the second testing phase, thus demonstrating that the gesture UI is very stable in terms of success rate. A1 (Previous) and A13 (Like) reach the maximum rate of 100% throughout the experiment and their symmetric task, *i.e.,* A2 (Next) and A14 (Dislike) are pretty close. These gestures are the most familiar to the participants, so they were easy to discover and repeat. Only two atomic tasks have a success rate below 80% in the learning phase: A4 (Disable full screen) and A9 (Rotate clockwise). In the first testing phase, only task A9 has a success rate below 80%, while in the second testing phase, the overall success rate decreases slightly and three tasks have a success rate lower than 80%: tasks A9, A10 (Rotate anticlockwise) and A16 (Rewind). The low success rate of the rotating (anti-)clockwise tasks can be explained by the high sensitivity of the system to the users' hand pose. For example, if the users' fingers were too far apart when performing a "grab" gesture, it was not recognized by the system. To a lesser extent, this high sensitivity may have affected the recognition of gestures involved in other tasks, such as A15 (fast forward), A16 (rewind), and A17 (pan). Regarding compound tasks, C1 (Photo) and C2 (Video) had a higher success rate than task C3 (Map) in both the first and second testing phases. The Pan gesture used in the Maps application was not elicited, which may have impacted the application's UX. In addition, this task required more precision than the other two, as it required participants to accurately manipulate the map so that two markers were placed on each side of the screen.

Table 4
Number of Gesture Trials for successful completion of each atomic task.

Id	Task	Learning phase		Testing phase 1		Testing phase 2	
		Average	Median	Average	Median	Average	Median
A1	Previous	1.2	1	1.6	1	1.7	1
A2	Next	1.1	1	1.3	1	1.6	2
A3	Enable full screen	1.5	1	1.5	1	1.1	1
A4	Disable full screen	1.2	1	1.7	2	1.5	1
A5	Zoom in	1.4	1	1.3	1	1.5	1
A6	Zoom out	1.4	1	1.1	1	1.2	1
A7	Volume up	1.3	1	1.3	1	1.1	1
A8	Volume down	2.0	2	1.5	1	1.3	1
A9	Rotate clockwise	2.1	1.5	1.6	1.5	1.9	1.5
A10	Rotate anti-clockwise	1.8	1.5	2.1	1.5	2.4	2.5
A11	Play	1.3	1	2.0	1	1.6	1
A12	Pause	1.4	1	1.5	1	1.8	1.5
A13	Like	1.0	1	1.2	1	1.3	1
A14	Dislike	1.0	1	1.1	1	1.0	1
A15	Fast-forward	1.7	1	1.7	1	1.9	2
A16	Rewind	1.4	1	1.7	1	1.4	1
A17	Pan	1.3	1	1.3	1	1.5	1

5.1.2. Number of Gesture Trials

Table 4 shows the average and median of Number of Gesture Trials needed to successfully complete the atomic tasks. These measures are presented for the learning session and the two testing phases, allowing in particular the study of the memorability of gestures over time. Overall, most tasks were completed with a single trial and constantly across phases, apart from the rotations, clockwise or anticlockwise. More trials were needed to pass the tasks in testing phase 2, especially for A1 (previous), A2 (next), A12 (pause) and A15 (fast forward). In contrast, the tasks of increasing or decreasing the volume (A7 and A8) required fewer trials over time.

5.1.3. Gesture Task Completion Time

Fig. 4 shows the gesture task average time for the seventeen atomic tasks across the three phases. Overall, navigation gesture tasks, such as A1 (Previous) and A2 (Next), as well as familiar tasks, such as A11 (Play), A12 (Pause), A13 (Like), and A14 (Dislike), are generally the fastest, as they are associated with simple commands that are usually familiar to participants. Most atomic tasks involving manipulation gestures, such as A5 to A10 and A15 to A17, took longer to complete than other simple tasks. There were more small variations in the way participants performed these gestures (*e.g.,* slightly different hand poses, such as folding all other fingers while pinching out with index and thumb vs. keeping all other fingers unfolded for zooming in), which may have made it more difficult to remember and perform the exact gesture expected by the system. The average completion time for some symmetric tasks, such as between tasks A9 (Rotate clockwise) and A10 (Rotate anti-clockwise), or between tasks A5 (Zoom in) and A6 (Zoom out) are very different, perhaps because some directions of a gesture are more intuitive than others, like rotating the wrist in a clockwise direction is considered familiar to right-handed people. Confusion can arise when the participant interprets a directional gesture in the direction of the target (*e.g.,* Next induces a gesture to the right), or when he interprets the gesture as erasing the source (*e.g.,* Next induces a gesture to the left).

While we might expect a decrease in time over the course of the phases due to the learning effect, we note that the average completion time does not necessarily decrease for each gesture task. On the contrary, all three variations can be observed: *decrease* for A5 (Zoom in), A6 (Zoom out), A7 (Volume up), A8 (Volume down), A10 (Rotate anticlockwise), A14 (Dislike), A16 (Rewind), A17 (Decrease) *increase* for A1 (Previous), A3 (Enable full screen), A4 (Disable full screen), A9 (Rotate clockwise), A11 (Play), or *relative stagnation* for A2 (Next), A12 (Pause), A13 (Like), and A15 (Fast forward).

To examine these variations more closely, we performed a series of Wilcoxon signed-rank tests for paired samples (since all participants performed the same tasks) to determine statistically significant differences between the times of these individual gesture tasks. These differences are reproduced in Fig. 4. For example, A4 (Disable full screen)

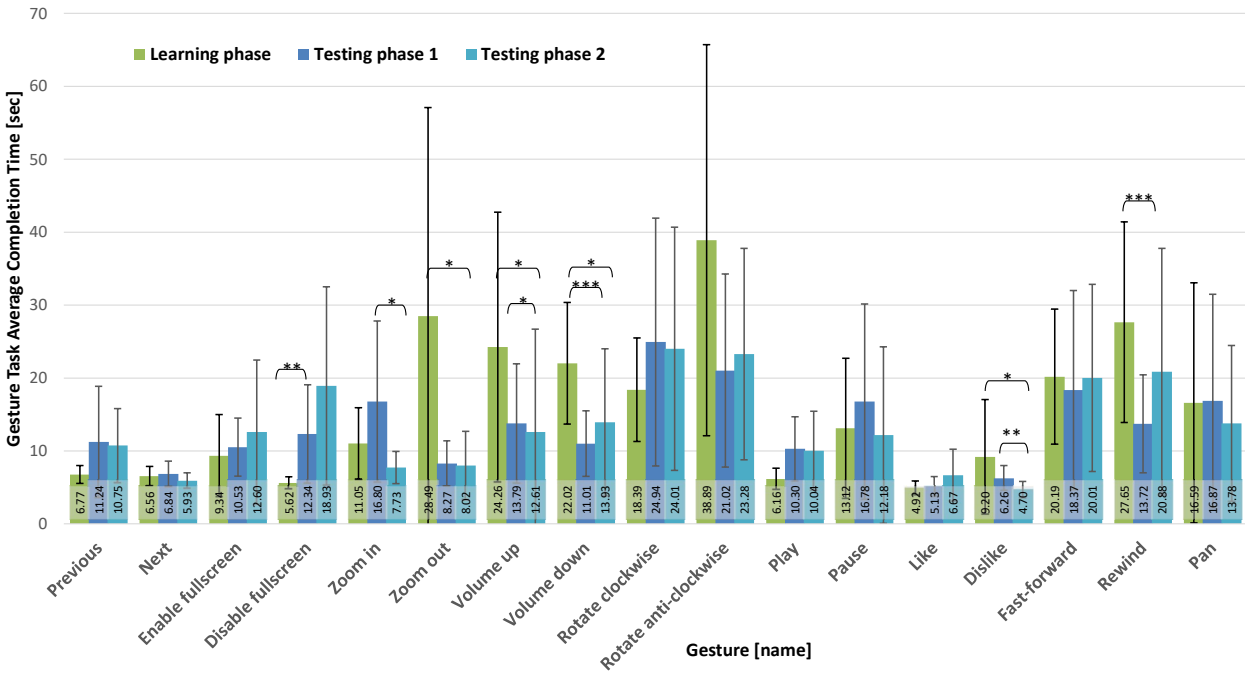


Figure 4: Gesture task average completion time for the seventeen atomic tasks. Error bars show a confidence interval of 95%. Significance levels result from a Wilcoxon signed-rank test for paired samples.

significantly increased from the learning phase to the testing phase 1 (M=76.5, z-$score$=2.04, p=.019**, r=.35). On the other hand, A7 (Volume up) significantly decreased from the learning phase to the testing phase 1 (M=76.5, z-$score$=2.98, p=.00067***, r=.51) and from this phase to testing phase 2 (M=76.5, z-$score$=1.65, p=.049*, r=.28), which suggests a real learning effect that converges towards a representative time.

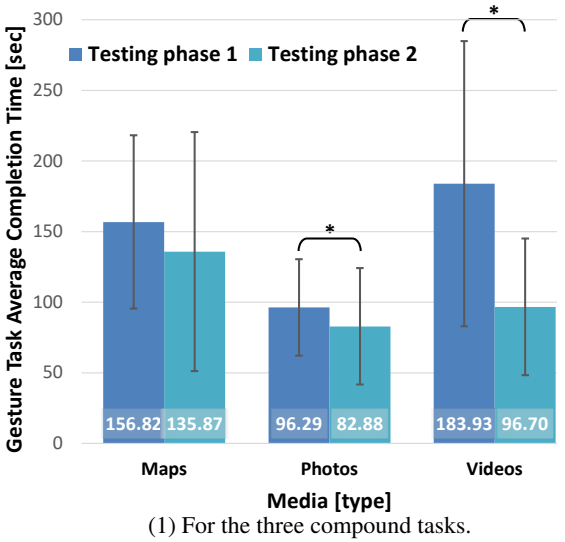
Fig. 5–1 shows the time for the three compound tasks per media type: for all three types, a time decrease is observed from testing phase 1 to testing phase 2, without any significance for Maps (M=76.5, z-$score$=1.56, p=.006, *n.s.*) and with some significance for Photos (M=76.5, z-$score$=1.66, p=.049*, r=.28) and for Videos (M=76.5, z-$score$=2.17, p=.013*, r=.37). Even better, when we calculate the average time for all task types in the three phases, we see a real reduction in the average time (Fig. 5–2): the learning phase definitely requires more time than the last testing phase (M=20952.5, z-$score$=3.17, p=.00071***, r=.13), which itself requires less time than the first testing phase (M=20520.5, z-$score$=1.92, p=.0261*, r=.08).

5.1.4. Gesture Recognition Rate

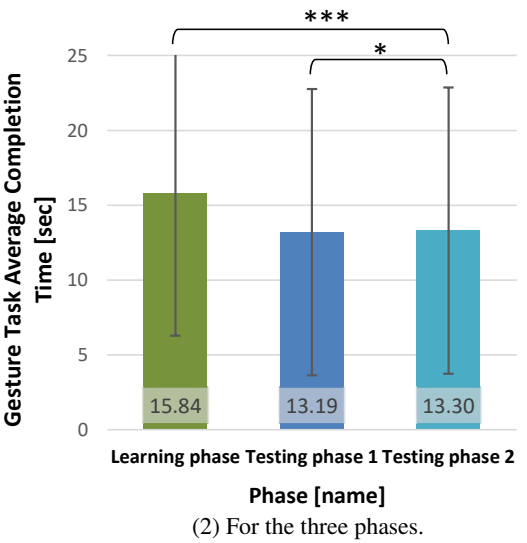
Fig. 6 shows the GESTURE RECOGNITION RATE for the seventeen gesture tasks subject to experiment sorted in decreasing order of their value. According to the definition, this rate represents the ratio between the successful gesture trials and the total amount of gesture trials. This rate, therefore, does not represent the system recognition rate, but an estimation of how many trials the participants needed to complete the gesture tasks on average. For example, the worst gesture task A10 (Rotate anti-clockwise) was recorded as totalizing 15 successful trials on a total amount of 48 trials, which gives $\frac{15}{48}$=31%. On the other hand, the A14 (Dislike) gesture task was always successful: participants issued a gesture that was adequate and correctly recognized.

5.2. Questionnaires and Interviews

In this second part, we are interested in the questionnaires, including the UEQ+ scales as well as the 3 new qualitative ad hoc scales, and the interviews. In particular, we show the interest of integrating this type of element into an interface evaluation, in addition to quantitative measurements.



(1) For the three compound tasks.



(2) For the three phases.

Figure 5: Gesture task average completion time. Error bars show a confidence interval of 95%. Significance levels result from a Wilcoxon signed-rank test for paired samples.

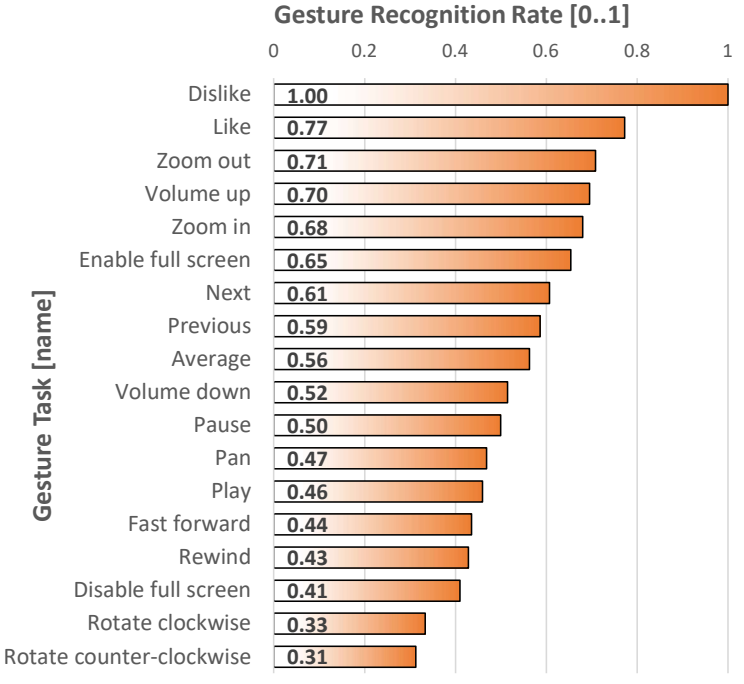


Figure 6: Gesture Recognition Rate for the seventeen gesture tasks.

5.2.1. UEQ+ Questionnaires

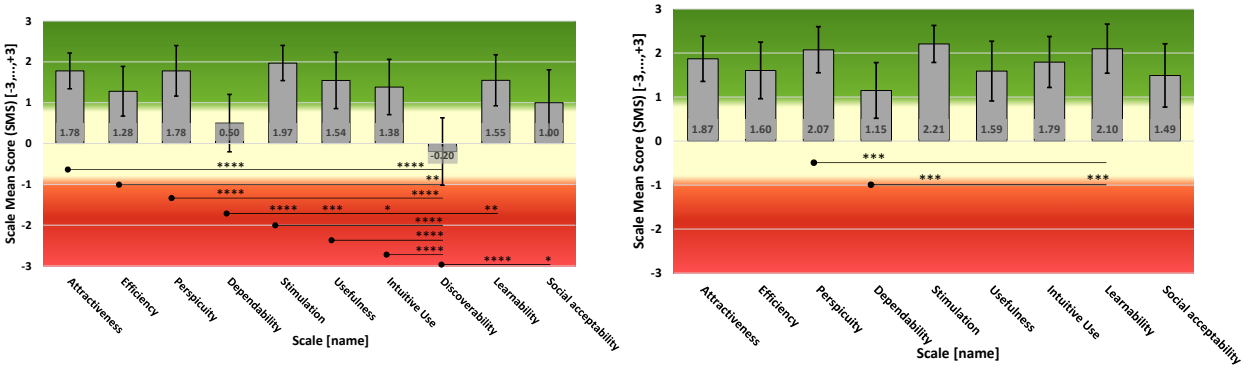
Participants' responses to the scales selected for this study were encoded in a spreadsheet to obtain the Scale Mean Scores for the first session (Fig. 7–1), for the second session (Fig. 7–2), and for their comparison (Fig. 9–1) as well as for their Scale Mean Importances (Fig. 9–2). Since all scales were 7-point Likert scales, the Scale Mean Score (SMS) for each scale belongs to the interval [-3,...,+3], where a score of -3 expresses the weakest value of the scale and +3 expresses the strongest value of the scale.

Table 5
Consistency (Cronbach's α: G=good, V=very good, E=excellent) and Interrater reliability (Kendall's W: V=very weak, W=weak, M=moderate, S=strong) of Scale Mean Scores (UEQ+ scales and new scales).

Scale	Cronbach's α (inter.)		Kendall's W (p-value, inter.)	
	S1	S2	S1	S2
Attractiveness	0.82 (V)	0.96 (E)	0.01 (0.92, V)	0.039 (0.59, W)
Efficiency	0.82 (V)	0.93 (E)	0.42 (<0.0001, S)	0.25 (0.0066, M)
Perspicuity	0.89 (V)	0.84 (V)	0.04 (0.61, V)	0.054 (0.45, W)
Dependability	0.85 (V)	0.92 (E)	0.08 (0.28, V)	0.21 (0.017, M)
Stimulation	0.81 (V)	0.93 (E)	0.09 (0.19, V)	0.010 (0.16, V)
Usefulness	0.92 (E)	0.97 (E)	0.17 (0.033, W)	0.048 (0.50, V)
Intuitive Use	0.77 (G)	0.89 (V)	0.04 (0.55, V)	0.041 (0.57, V)
Discoverability	0.90 (E)	–	0.03 (0.56, V)	–
Learnability	0.84 (V)	0.95 (E)	0.15 (0.08, W)	0.04 (0.53, W)
Social acceptability	0.74 (G)	0.95 (E)	0.08 (0.25, V)	0.02 (0.73, V)

Consistency and Reliability. Table 5 shows the reliability of the seven scales with regard to consistency and interrater reliability. Overall, all scales are well consistent ($\alpha \geq 0.70$) for the first session and very well consistent ($\alpha \geq 0.80$) for the second, and all values increased from the first session to the second. The inter-rater reliability ranges from very weak for five scales to strong for one of them (Efficiency). Little or no variations were observed after the second session. We ran a series of Kolmogorov-Smirnov tests to check the normality of these scales. None of them passed the normality test. For example, Attractiveness for both the first session (S1, *KS distance*=0.24) and the second session (S2, *KS distance*=0.28) did not pass the normality test (α=.05, p<.0001****). A series of Shapiro-Wilk tests confirmed the same results. For example, Attractiveness for both the first session (S1, W=0.87) and the second session (S2, W=0.79) did not pass the normality test (α=0.05, p<0.0001****). Consequently, we performed a series of Wilcoxon matched-pairs signed rank tests (two-tailed) for all scales within each session (S1, S2) and across sessions (S1 *vs.* S2) to investigate any potential difference among them.

Intra-session Scale Mean Scores (S1, S2). If we stick to the original interpretation that a scale mean score greater than or equal to 0.8 represents a positive evaluation (Schrepp and Thomaschewski, 2019), then all mean scores for the **first session** are considered positive (ranging from M=1.00 for Social acceptability and M=1.28 for Efficiency to M=1.97 for Stimulation), except Dependability (M=0.50, SD=2.22) and Discoverability (M= − 0.20, SD=3.08) considered as having a value in the neutral zone (depicted in yellow). These neutral values can be explained by the cognitive destabilization of participants producing these kinds of gestures for the first time in their life and being puzzled by them. However, this phenomenon is observed transiently insofar as the gestures, initially felt to be



(1) Session 1. (2) Session 2.

Figure 7: Scale Mean Scores for all scales (UEQ+ and new ones). Error bars show a confidence interval of 95%. Significance levels result from a Kruskal-Wallis test for independent samples.

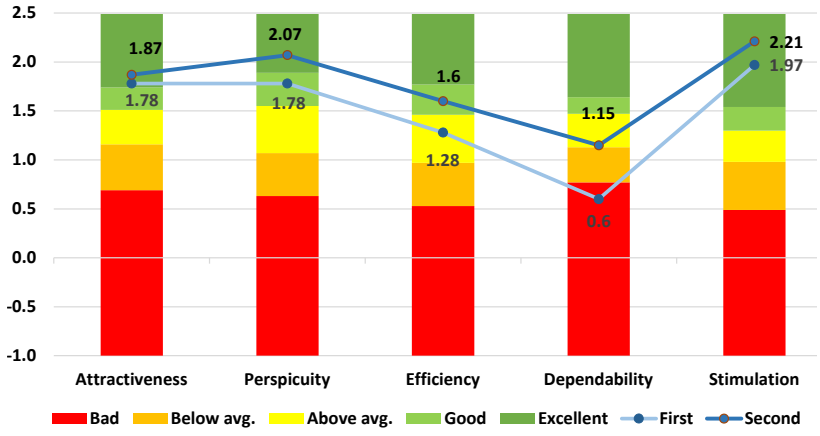


Figure 8: Benchmarking of scales for the two sessions.

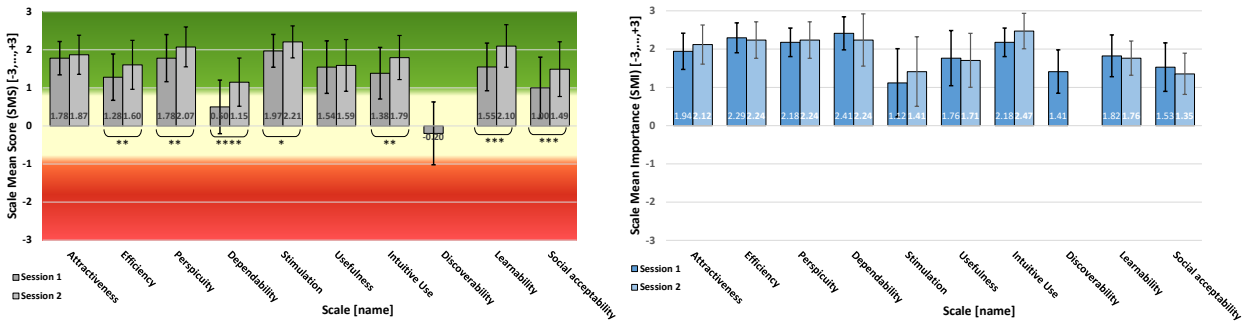
difficult to discover and produce because they are unknown, then become easier to produce. The negative score of DISCOVERABILITY is explained by the lack of instructions about potential gestures during the discovery phase, which was decided on the purpose and represents a real challenge. When Apple launched the first version of the iPhone, every commercial showed most of the nine iOS gestures (*i.e.,* Tap, Double tap, Touch and Hold, Drag-Flick, Pinch, Spread, Rotate, Two fingers drag-down, and Two fingers Drag-up) to teach customers to use the gesture interface since nobody has seen them before (Wigdor and Wixon, 2011).

If we interpret the mean scores according to the partial benchmarking (Table 1 in Schrepp, Hinderks and Thomaschewski (2017)), then ATTRACTIVENESS is assessed as "excellent" (M=1.78 $\geq$ 1.75), EFFICIENCY is assessed as "above average" (M=1.28 $\geq$ 0.98), PERSPICUITY is assessed as "good" (M=1.78 $\geq$ 1.56), DEPENDABILITY is assessed as "bad" (M=0.50 $<$ 0.78), and STIMULATION is assessed as "excellent" (M=1.97 $\geq$ 1.55), as shown in Fig. 8. A Kruskal-Wallis test for independent samples representing the ten scales with multiple comparisons among them was performed to identify the statistically significant differences shown in Fig. 7–1: for example, DISCOVERABILITY is significantly lower than the other gesture scales during the same session with respect to LEARNABILITY-$F=-329.0$, $p<0.0002^{****}$ and to SOCIAL ACCEPTABILITY-$F=-230.2$, $p<0.05^{*}$). Similarly, DEPENDABILITY is judged lower than most other scales, while ATTRACTIVENESS is judged better than DEPENDABILITY and than DISCOVERABILITY.

Regarding the **second session**, all scale mean scores are again above the threshold of 0.8 (Schrepp and Thomaschewski, 2019), ranging from a minimum for DEPENDABILITY (M=1.15, SD=1.33) to a maximum for STIMULATION (M=2.21, SD=0.79). The mean scores resulting from this second session have all improved according to the benchmark (Table 1 in Schrepp et al. (2017), see Fig. 8): ATTRACTIVENESS is assessed as "excellent" (M=1.87 $\geq$ 1.75), PERSPICUITY is assessed as "excellent" (M=2.07 $\geq$ 1.90), EFFICIENCY is assessed as "good" (M=1.60 $\geq$ 1.60), DEPENDABILITY is assessed as "above average" (M=1.15 $\geq$ 1.14), and STIMULATION is assessed as "excellent" (M=2.21 $\geq$ 1.55). Although DEPENDABILITY increased since the first session, the Kruskal-Wallis test with multiple comparisons reveals that DEPENDABILITY is again significantly lower than STIMULATION and LEARNABILITY (Fig. 7–2). Since DISCOVERABILITY is relevant only for the first session, it no longer appears in this second session.

Inter-session Scale Mean Scores (S1 and S2). Although DEPENDABILITY knew a neutral score during the first session, it experienced a positive increase of 130% (from M=0.5 to M=1.15, SD=1.80) in the second session, which leads to a positive evaluation. The mean scores for S2 are all higher than those obtained for S1 and range from M=1.15 for DEPENDABILITY to M=2.21 for STIMULATION. The mean score for some scales increased even a lot, such as EFFICIENCY (from M=1.28 to M=1.60), PERSPICUITY (from M=1.78 to M=2.07), and INTUITIVE USE (from M=1.38 to M=1.79). More important is the increase in LEARNABILITY starting from M=1.55 to M=2.10, the second highest score of all scales.

Based on a two-tailed Wilcoxon signed-ranked test, we found a statistically significant increase in the scale mean score for seven scales among the nine used, which denotes real progress across sessions. Such a difference was found for EFFICIENCY (W=332, p=.0089**), thus suggesting that participants felt more efficient in producing the gestures during



(1) Scale Mean Scores.

(2) Scale Mean Importances.

Figure 9: Comparison of Scale Mean Scores and Scale Mean Importances for all scales (UEQ+ and new ones) across sessions (Session 1=left in dark grey, Session 2=right in light grey). Error bars show a confidence interval of 95%. Significance levels result from a Wilcoxon signed-rank test for paired samples.

the second session than during the first one. We also found a strong positive correlation (r_s=0.62, p<0.0001****): the more a participant passes a new session, the more efficient it becomes. If we had conducted still other sessions, it is likely to observe an increase in EFFICIENCY until reaching an asymptote. Similarly, a significant increase was found for PERSPICUITY (W=263, p=.0094**), suggesting that participants became more familiar with the gesture UI during the second session than during the first. Same for DEPENDABILITY, STIMULATION, INTUITIVE USE, LEARNABILITY, and SOCIAL ACCEPTABILITY, thus suggesting that the participants felt overall much more comfortable using gesture UI in the long-term case than in the short term.

No significant differences were found for ATTRACTIVENESS (W=114, p=.49, *n.s.*) and USEFULNESS (W=52, p=.62, *n.s.*). All in all, we see the benefits of a positive learning effect: during the first session, some participants felt unsafe or unsure of how to correctly produce gestures mainly because they had never done it before. After the second session, they felt more safe and confident about reproducing and reusing the gestures they learned. A similar phenomenon was observed when 2D stroke gestures were introduced for the first time on the Apple iPhone: People were unsure about what gestures to do when using the interface, but when they realized them, they felt safer and more able to reproduce the gestures quite efficiently (Wigdor and Wixon, 2011).

The KPI (Key Performance Indicator) computed by UEQ+ (Hinderks, Winter, Schrepp and Thomaschewski, 2019) started from M=1.44 (SD=0.80) for the first session to M=1.74 (SD=0.94) for the second, which confirms the overall positive evolution.

Intra-session Scale Mean Importance (S1, S2). Regarding the **first session**, all importance scores are very well consistent: the global Cronbach's α=0.85 $\geq$ 0.80, all individual α are above 0.80, except DEPENDABILITY (α=0.79 $\leq$ 0.80). However, the inter-rater reliability is very weak (Kendall and Babington Smith (1939)'s W=0.081, p=0.21). All importance ratings are significantly higher than the median value Md=4, ranging from STIMULATION (W=74, p=0.033*) to ATTRACTIVENESS (W=148, p<0.0001****). The STIMULATION (M=5.1) was rated significantly (W=−25, p=0.046*) less important than PERSPICUITY (M=6.2) and significantly (W=−55, p=0.0020**) less important than DEPENDABILITY (M=6.4). The USEFULNESS was rated significantly (W=−28, p=0.015*) less important than DEPENDABILITY. The INTUITIVE USE was rated significantly more important (W=25, p=0.046*) than STIMULATION (Fig. 9–2). Regarding the **second session**, all importance scores are very consistent: the global Cronbach's α=0.85 $\geq$ 0.80, all individual α are above 0.80, ranging from USEFULNESS (α=0.80) to INTUITIVE USE (α=0.85). The interrater reliability is again very weak (Kendall's W=0.057, p=0.44). All importance ratings are significantly higher than the median value Md=4, ranging from STIMULATION (W=84, p=0.014*) to ATTRACTIVENESS (W=148, p<0.0001****). A Friedman test with multiple comparisons reveals that the importance was not rated differently depending on the scale (F=10.69, p=0.098, *n.s.*).

Inter-session Scale Mean Importance (S1 and S2). No significant difference was found between the importance of scales (Fig. 9–2) across sessions (all Wilcoxon tests returned a p>0.05), thus suggesting that participants did not attach more or less importance to the scales during the second session than during the first.

5.2.2. Discoverability, Learnability, and Social Acceptability

Three new qualitative scales, *i.e.,* DISCOVERABILITY, LEARNABILITY, and SOCIAL ACCEPTABILITY, were added to extend the evaluation beyond the current scope and to cover gesture interaction. The three last rows of Table 5 show that their Cronbach coefficient α was excellent from the beginning of DISCOVERABILITY (α=.90) until the end of LEARNABILITY and SOCIAL ACCEPTABILITY (α=.95), suggesting that the participants responded consistently and are reliable (Table 5). The coefficient of reliability is approximately the same for all scales at the end of the second session. The values of Kendall's coefficient of concordance W are interpreted as "weak" or "very weak", indicating that there was no particular agreement among the participants in the way they assessed the three scales (Table 5). This is pretty much justified by the various types of profile: participants having very diverse profiles have assessed the gesture UI in their own way, which is not necessarily in agreement with other participants.

Fig. 10 shows how the participant's answers to the items of these three new scales are distributed. Since the distribution of the answers does not follow a normal distribution (all Kolmogorov-Smirnov normality tests did not pass with α=0.05), we calculated a series of Wilcoxon signed rank tests for a single sample with ties and continuity correction (with 10,000 iterations) to determine whether the items would significantly depart from the median value Md=4, either above or below. None of the three elements of DISCOVERABILITY were significantly different from the median: D_1=' 'I could easily guess how to use the interface" (M=3.76, SD=1.83, Mdn=4, p=.23, *n.s.*), D_2="It was not complicated to find out how to control the interface" (M=3.59, SD=1.78, Mdn=3, p=.18, *n.s.*), and D_3="The gestures to control the interface were easy to discover" (M=4.06, SD=1.55, Mdn=5, p=.47, *n.s.*).

All other items of LEARNABILITY and SOCIAL ACCEPTABILITY were significantly above the median during both sessions, except for the item SA_2 (M=4.71, SD=1.74, Mdn=5, p=.067, *n.s.*). For example, L_1 (M=5.88, SD= 0.90, Mdn=6) is significantly ($p \leq .001^{***}$) above the median with a large effect size (r=.86). We believe that DISCOVERABILITY was rated low due to the new interaction modality that participants were not used to, which induced a disruptive nature and cognitive destabilization in the beginning. We captured this scale only once since gestures need to be discovered the first time participants are confronted. This scale is likely to be rated differently during a subsequent session. Fortunately, both LEARNABILITY and SOCIAL ACCEPTABILITY were positively assessed by the participants: not only the items were almost very positively assessed but significantly better during the second session than during the first session.

We finally computed another series of Wilcoxon signed-rank tests to investigate whether participants answered in a different way to the items during the second session compared to the first. All six items were assessed significantly better during S2 than during S1, which denotes a real progression of the answers, even after only one session. For instance, the third item of SOCIAL ACCEPTABILITY SA_3 during S1 (M=5.24) is different than during S2 (M=5.53) in a significant way (*z-score*=3.62, p=.00014***) with a large effect size (r=.87), thus indicating that participants felt significantly less embarrassed with the gesture UI during the second session than during the first one.

5.2.3. Interviews

The use of interviews allows us to confirm certain results but also to reveal others, given the more free and open nature of the questions asked. Regarding the advantages of this type of interface, intuitiveness was mentioned by 9 of the 17 participants positively.

"If I compare it to a computer, it is true that it is just gestures and it's easier to remember than shortcuts. That I think is a bit the basis of the gesture." (P3, male, 23).

Other arguments in favor of this type of interface are its innovative and original side, as well as its usefulness and practicality perceived as superior. Finally, the fun provided by using this system is also an argument that emerged for some participants. It is of particular interest, as it is considered one of the fundamental pillars of providing a rich consumer experience (Holbrook and Hirschman, 1982). Regarding the first negative point that came to the minds of the participants, the sensitivity of the system, linked to the recognition performance, emerged in the interview of 15 of the 17 participants. This problem can even lead to frustration and is often linked to the user's feeling that he may be the cause of the problem, speaking for example of poor control on his part.

"It is frustrating when it confuses gestures. For example, when I want to turn up the volume and it changes my video, it's a bit scary." (P6, woman, 22)

Another problem that emerged in 10 interviews is the direction of movement forward or backward in the video, considered to be reversed. It should also be noted that the 7 people who did not mention this in their interview just did not bring this subject up on their own. Therefore, this does not mean that they consider the gesture to be correct in the way that it is currently implemented. The fact of making a move to the right to go back in the video and to the left to go

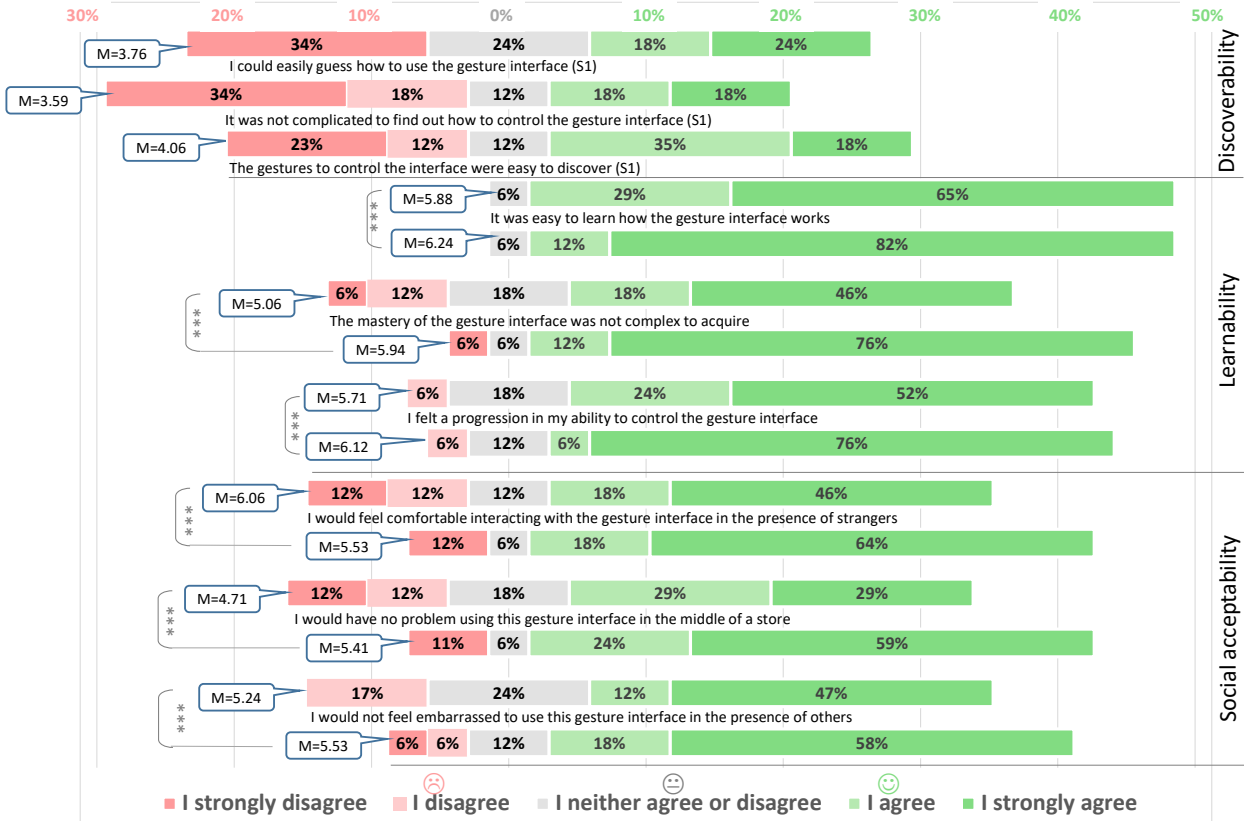


Figure 10: Distribution of participant's answers to the three new scales: DISCOVERABILITY (only during the first session S_1), LEARNABILITY (for both sessions S_1 and S_2, and SOCIAL ACCEPTABILITY (for both sessions as well) (M=mean, p<.001***).

forward was chosen during the gesture elicitation study (GES) (Wobbrock et al., 2009). However, now that the gesture has been implemented in a real application, the participants feel like they are grabbing the cursor and moving it. In this logic, it would be necessary to move to the right to move forward and to the left to move back. This is an example of the difference that can be created between a gesture elicitation study (see Villarreal-Narvaez, Sluÿters, Vanderdonckt and Vatavu (2024) for a review) and a real implementation. It thus seems necessary to carry out a systematic evaluation of each system, even if it was built by users, and also to leave the possibility to the participant to freely approach any subject that he wishes. It also shows the advantage of organizing interviews, as this problem would have been invisible if we had focused only on the other evaluation methods.

Finally, some other negative points were cited to a lesser extent. Three older participants found the gesture interaction to be unintuitive and took a long time to get used to. Another point is muscle fatigue, mentioned by 4 participants, and finally, problems understanding how the interface works. This last problem is however related to this data collection since we initially left the participants to discover for themselves without any help. To ensure a correct understanding and use of this type of interface, it is therefore necessary to be particularly attentive to the learning of users, for example, via a video tutorial and even personal support.

5.3. Added Value of Adding and Integrating such Approaches

This concrete evaluation shows the advantages of this method, first in combining different types of measures to have complementary views of the same result observed using different methods. On the one hand, the task success rate remains relatively high and the median success is mostly one, which attests to the quality of the interface. This result is confirmed by our questionnaires, where the majority of UEQ+ scales are considered excellent or good. However, we also observe a divergence in these data. In atomic tasks, the situation is slightly better in the first session rather

than the second. However, when we look at the questionnaires, the experience is considered to be better in the second session. So even though users on average had more difficulty a week later, they were more comfortable and satisfied with the interface. This is an example of rich information that can be highlighted with this type of method. By relying on the quantitative approach, we would not have had this information about the overall experience, which could have led to distorted decision making based solely on the recognition of gestures rather than the overall user experience.

Another advantage of this interface is the division between atomic and compound tasks. Focusing on the gestures themselves, we could see the weakness of certain movements. In this case, the rotation movements appeared weaker in the atomic tasks but was not apparent in the compound task of the photos manipulation. On the other hand, the Maps application stands out as significantly weaker, whether by observing the success rate, recognition rate or the time needed to complete the task. This problem would have been completely invisible if we had only looked at atomic tasks, staying at the gesture level without considering actual use cases.

Finally, debriefing interviews make it possible to confirm the overall quality of the interface, particularly in terms of intuitiveness, which also emerges from the questionnaires. Apart from that, they allowed us to highlight a problem unknown to the experimenters, namely the direction of movement to make to speed up the video or go back. This problem would not have been discovered if we had been satisfied with our objective measures and questionnaires. This is also an example confirming the need to carry out user tests of the application even if it is based on a gesture elicitation study, as it was for the interface used in this case study.

The results obtained through our combination of measures suggest that there is no optimal solution that optimizes all measures at once. Rather, a trade-off should be decided between these measures. To support this type of decision, we can complement our analysis by performing an Importance-Performance Analysis (IPA) (Martilla and James, 1977), which determines the satisfaction of the participants by querying their performance with respect to their preference for a set of measures and presenting it in a plot. The recommendations for action are derived from the arrangement of the menus in the plot, which consists of four quadrants (Martilla and James, 1977; Magal and Levenburg, 2005) adapted as follows (Fig. 11):

1. *Quadrant 1* (Q1, colored in green) corresponds to high-performance values and high-preference values for gesture measures. It is labeled "Keep up the good work" because the measures contained in this quadrant are positively assessed by the participants, who acknowledge both their preference (*e.g.,* in terms of GESTURE AGREEMENT RATE on the Y axis) and their performance (*e.g.,* in terms of GESTURE RECOGNITION RATE on the X axis). Designers are encouraged to maintain current strategies for these measures and retain these gesture tasks as ideal candidates. For example, the best candidates for the GESTURE RECOGNITION RATE are A14 (Dislike), A13 (Like), A2 (Next), A1 (Previous), and A7 (Volume up), because they exhibit a value above the average for performance and a value above the average for reference.
2. *Quadrant 2* (Q2, colored yellow) corresponds to high performance values and low preference values for gesture measures. It is labeled "Concentrate here" because the gestures that fall into this quadrant should receive the highest priority. Action should be taken to address the challenges raised by these gesture tasks because the participants performed well with them but did not really agree to use them for their own reasons. For example, participants do not like to be forced to operate efficiently in a continuous way. For example, these cases for the GESTURE TASK SUCCESS RATE include A3 (Enable full screen), A11 (Play), A4 (Disable full screen), A5 (Zoom in), and A8 (Volume down).
3. *Quadrant 3* (Q3, colored orange) corresponds to low performance values and low preference values for gesture measures. It is labeled "Low priority" because gestures located in this quadrant are estimated not essential and not performing by the participants and can be discontinued in the design process without detriment.
4. *Quadrant 4* (Q4, colored in ivory) corresponds to low-performance values and high-preference values for gesture measures. It is labeled "Possible overkill" because participants do not perform well with the gestures located in this quadrant, but perceive them as highly agreed upon.

6. Conclusion and Future Work

This paper defines a method for explicitly evaluating gesture user interfaces, dividing the analysis by gesture, use case, and overall subjective experience. This method combines quantitative measures, particularly specific to memorability, qualitative scales, and interviews. In particular, we have defined three new qualitative scales that are particularly relevant in the context of gestural interaction, namely, learnability, discoverability, and social acceptability.

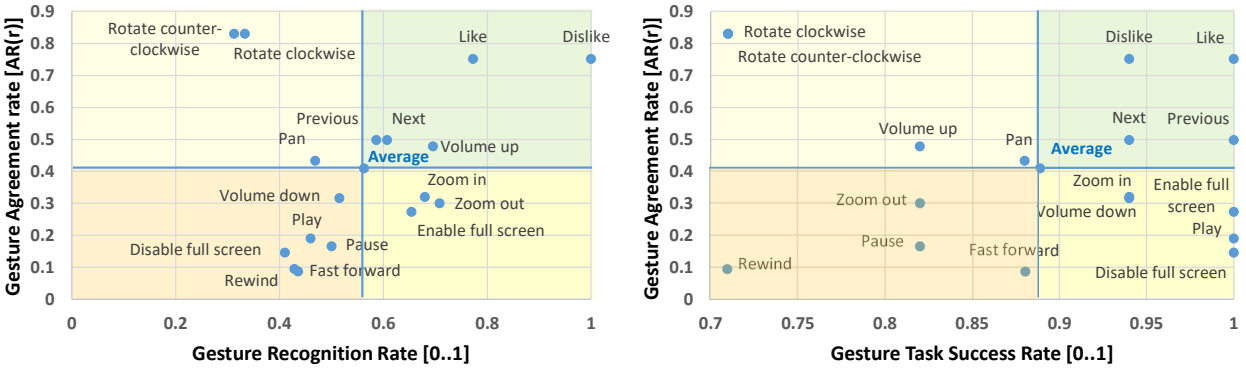


Figure 11: Importance-Performance Analysis of two gesture measures with respect to their Gesture Agreement Rate: Gesture Recognition Rate (left), Gesture Task Success Rate (right).

We applied and validated this method with Lui, a mid-air gesture user interface for browsing multimedia content by a general audience. We then discussed our results and provided some key takeaways.

The added scales provide a more comprehensive assessment of gestural interfaces by considering subjective user experiences and perceptions that are critical to evaluating the quality of the interface. Second, we propose a novel approach to gestural interface evaluation that combines quantitative measures, qualitative scales, and debriefing interviews to obtain a holistic view of the experience. This approach recognizes the importance of both objective and subjective evaluation methods. We highlight the importance of multi-session evaluation, particularly for measuring the learnability of gestural interfaces. This is important as gestural interfaces are often more complex than traditional graphical user interfaces and require users to learn new interaction paradigms. By obtaining a more complete understanding of the experience, designers and developers can make informed decisions about how to improve their interfaces and make them more user-friendly. Another advantage of this method is the division between atomic and compound tasks. In the case study, focusing on the gestures themselves, we could see the weakness of the rotation movements, which was not apparent in the compound task of the photos. On the other hand, the Maps application stands out as clearly weaker, which would have been invisible by just looking at the atomic tasks. Finally, the debriefing interviews allowed us to confirm the overall quality of the interface, particularly in terms of intuitiveness, which also emerged from the questionnaires. Apart from that, they allowed us to highlight a problem unknown to the experimenters, which would not have been discovered if we had been satisfied with our objective measurements and the questionnaires.

This work also presents limitations and avenues for future work. The first important point is the use of our three new scales, which have been chosen because of their good validity results in other studies. It would, however, be interesting to change their format in order to match the one used in UEQ+, therefore functioning as an extension of the latter. To increase the ecological validity of this evaluation method, it should be applied to other interfaces and contexts. This would also allow us to refine the method, for instance in terms of the scales used. Another future research would be to adapt this method to other contexts and interaction mediums, in particular for multisensory input. It would also be an opportunity to increase the ecological validity of this method and popularize its use.

As main contributions, we highlight the importance of having a multi-session evaluation approach to measure the learnability of gestural interfaces. This approach involves testing users on multiple occasions, which provides a more accurate measure of long-term experience with the interface. This is especially important as gestural interfaces are often more complex than traditional graphical user interfaces and require users to learn new interaction paradigms. We also introduce three new scales to complete the UEQ+ evaluation method, which were specifically designed to measure the learnability, discoverability, and social acceptability of gestural interfaces. These scales provide a more comprehensive assessment of gestural interfaces by including subjective user experiences and perceptions that are critical to evaluating the quality of the interface. Finally, we demonstrate the usefulness of debriefing interviews as a complement to questionnaires and objective measures. Debriefing interviews can help identify interface issues that may not be apparent from other evaluation methods and can provide researchers with valuable feedback to improve the interface.

References

Archer, D., 1997. Unspoken diversity: Cultural differences in gestures. Qualitative Sociology 20, 79–105. URL: https://doi.org/10.1023/A:1024716331692, doi:10.1023/A:1024716331692.

Bachmann, D., Weichert, F., Rinkenauer, G., 2018. Review of three-dimensional human-computer interaction with focus on the leap motion controller. Sensors 18, 2194. URL: https://doi.org/10.3390/s18072194, doi:10.3390/s18072194.

Barclay, K., Wei, D., Lutteroth, C., Sheehan, R., 2011. A quantitative quality model for gesture based user interfaces, in: Proceedings of the 23rd Australian Computer-Human Interaction Conference, Association for Computing Machinery, New York, NY, USA. p. 31–39. URL: https://doi.org/10.1145/2071536.2071540, doi:10.1145/2071536.2071540.

van Beurden, M.H.P.H., Ijsselsteijn, W.A., de Kort, Y.A.W., 2012. User experience of gesture based interfaces: A comparison with traditional interaction methods on pragmatic and hedonic qualities, in: Efthimiou, E., Kouroupetroglou, G., Fotinea, S.E. (Eds.), Gesture and Sign Language in Human-Computer Interaction and Embodied Communication, Proc. of Gesture and Sign Language in Human-Computer Interaction and Embodied Communication (GW '11), Springer, Berlin, Heidelberg. pp. 36–47. URL: https://link.springer.com/chapter/10.1007/978-3-642-34182-3_4, doi:10.1007/978-3-642-34182-3_4.

Bhuiyan, M., Picking, R., 2011. A gesture controlled user interface for inclusive design and evaluative study of its usability. Journal of Software Engineering and Applications 4, 513–521. URL: https://doi.org/10.4236/jsea.2011.49059, doi:10.4236/jsea.2011.49059.

Boyatzis, R.E., 1998. Transforming Qualitative Information: Thematic Analysis and Code Development. SAGE. ZSCC: 0017594 Google-Books-ID: _rfClWRhIKAC.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3, 77–101. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa, doi:10.1191/1478088706qp063oa. number: 2 ZSCC: 0099387 Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

Brooke, J., 1996. SUS-A 'Quick and Dirty' Usability Scale. CRC Press, London, UK. chapter 6. pp. 189–194. URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke, doi:https://doi.org/10.1201/9781498710411.

Calvary, G., Coutaz, J., Thevenin, D., Limbourg, Q., Bouillon, L., Vanderdonckt, J., 2003. A Unifying Reference Framework for multi-target user interfaces. Interacting with Computers 15, 289–308. URL: https://doi.org/10.1016/S0953-5438(03)00010-9, doi:10.1016/S0953-5438(03)00010-9.

Chuan, N.K., Sivaji, A., Ahmad, W.F.W., 2014. Proposed usability heuristics for testing gestural interaction, in: Proceedings of 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, IEEE Computer Society Press, Los Alamitos. pp. 233–238. URL: https://ieeexplore.ieee.org/document/7351840, doi:10.1109/ICAIET.2014.46.

Chuan, N.K., Sivaji, A., Ahmad, W.F.W., 2015. Usability heuristics for heuristic evaluation of gestural interaction in hci, in: Marcus, A. (Ed.), Design, User Experience, and Usability: Design Discourse, Springer International Publishing, Cham. pp. 138–148. URL: https://link.springer.com/chapter/10.1007/978-3-319-20886-2_14, doi:10.1007/978-3-319-20886-2_14.

Cockburn, A., Gutwin, C., Scarr, J., Malacria, S., 2014. Supporting novice to expert transitions in user interfaces. ACM Comput. Surv. 47. URL: https://doi.org/10.1145/2659796, doi:10.1145/2659796.

Dix, A., Finlay, J., Abowd, G., Beale, R., 2004. Evaluation techniques. Human-computer interaction .

Farhadi-Niaki, F., Etemad, S.A., Arya, A., 2013. Design and usability analysis of gesture-based control for common desktop tasks, in: Kurosu, M. (Ed.), Proceedings of International Conference on Human-Computer Interaction HCI '13, Human-Computer Interaction. Interaction Modalities and Techniques, Springer, Berlin, Heidelberg. pp. 215–224. doi:10.1007/978-3-642-39330-3_23.

Fonseca Brandao, A., Casseb, R., Almeida, S., Assis, G., Camargo, A., Min, L.L., Castellano, G., 2019. Investigation of fmri protocol for evaluation of gestural interaction applied to upper-limb motor improvement. SBC Journal on Interactive Systems 10. URL: https://seer.ufrgs.br/index.php/jis/article/view/84779.

Graichen, L., Graichen, M., Krems, J.F., 2019. Evaluation of gesture-based in-vehicle interaction: User experience and the potential to reduce driver distraction. Human Factors 61, 774–792. URL: https://doi.org/10.1177/0018720818824253, doi:10.1177/0018720818824253, arXiv:https://doi.org/10.1177/0018720818824253. pMID: 30694705.

Grossman, T., Fitzmaurice, G., Attar, R., 2009. A survey of software learnability: metrics, methodologies and guidelines, in: Proceedings of the ACM Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 649–658. URL: https://doi.org/10.1145/1518701.1518803, doi:10.1145/1518701.1518803.

Guerino, G.C., Valentim, N.M.C., 2020. Usability and user experience evaluation of natural user interfaces: a systematic mapping study. IET Software 14, 451–467. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-sen.2020.0051, doi:https://doi.org/10.1049/iet-sen.2020.0051, arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-sen.2020.0051.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, in: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. North-Holland. volume 52 of *Advances in Psychology*, pp. 139 – 183. URL: http://www.sciencedirect.com/science/article/pii/S0166411508623869, doi:https://doi.org/10.1016/S0166-4115(08)62386-9.

Hassenzahl, M., 2008. The interplay of beauty, goodness, and usability in interactive products. Hum.-Comput. Interact. 19, 319–349. URL: https://doi.org/10.1207/s15327051hci1904_2, doi:10.1207/s15327051hci1904_2.

Hinderks, A., Winter, D., Schrepp, M., Thomaschewski, J., 2019. Applicability of user experience and usability questionnaires. J. Univers. Comput. Sci. 25, 1717–1735. URL: http://www.jucs.org/jucs_25_13/applicability_of_user_experience.

Holbrook, M.B., Hirschman, E.C., 1982. The Experiential Aspects of Consumption Consumer Fantasies, Feelings, and Fun. Journal of Consumer Research 9, 132–140. doi:10.1086/208906. number: 2 ZSCC: 0009838.

ISO, 2018. ISO/IEC 9241, Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. standard. International Standard Organization. Geneva. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:en.

ISO, 2019. ISO/IEC 25010 - Software Quality Product Standard. standard. International Standard Organization. Geneva. URL: https://iso25000.com/index.php/en/iso-25000-standards/iso-25010?limit=3&limitstart=0.

Kendall, M., Babington Smith, B., 1939. The Problem of m Rankings. Annals of Math. Statistics 10, 275–287. URL: http://www.jstor.org/stable/2235668.

Lewis, J.R., Sauro, J., 2009. The factor structure of the system usability scale, in: Kurosu, M. (Ed.), Human Centered Design, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 94–103.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 22, 55–. URL: http://psycnet.apa.org/record/1933-01885-001.

Mackamul, E., 2022. Improving the discoverability of interactions in interactive systems, in: Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. pp. 1–5. URL: https://doi.org/10.1145/3491101.3503813, doi:10.1145/3491101.3503813.

Magal, S.R., Levenburg, N.M., 2005. Using importance-performance analysis to evaluate e-business strategies among small businesses, in: Proceedings of 38th Hawaii International Conference on System Sciences, IEEE Computer Society. pp. 176a–176a. URL: https://doi.org/10.1109/HICSS.2005.661, doi:10.1109/HICSS.2005.661.

Martilla, J.A., James, J.C., 1977. Importance-performance analysis. Journal of Marketing 41, 77–79. URL: https://doi.org/10.1177/002224297704100112, doi:10.1177/002224297704100112, arXiv:https://doi.org/10.1177/002224297704100112.

Montero, C.S., Alexander, J., Marshall, M.T., Subramanian, S., 2010. Would you do that? understanding social acceptance of gestural interfaces, in: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, Association for Computing Machinery, New York, NY, USA. p. 275–278. URL: https://doi.org/10.1145/1851600.1851647, doi:10.1145/1851600.1851647.

Nacenta, M.A., Kamber, Y., Qiang, Y., Kristensson, P.O., 2013. Memorability of pre-designed and user-defined gesture sets, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1099–1108. URL: https://doi.org/10.1145/2470654.2466142, doi:10.1145/2470654.2466142.

Neca, J., Duarte, C., 2011. Evaluation of gestural interaction with and without voice commands, in: Proceedings of IADIS International Conference Interfaces and Human Computer Interaction, pp. 69–76. URL: http://www.iadisportal.org/digital-library/evaluation-of-gestural-interaction-with-and-without-voice-commands.

Nielsen, J., 1994. Usability Engineering. Interactive Technologies, Elsevier Science. URL: https://books.google.be/books?id=95As2OF67f0C.

Norman, D.A., Nielsen, J., 2010. Gestural interfaces: A step backward in usability. Interactions 17, 46–49. URL: https://doi.org/10.1145/1836216.1836228, doi:10.1145/1836216.1836228.

Parthiban, V., Maes, P., Sellier, Q., Sluÿters, A., Vanderdonckt, J., 2022. Gestural-vocal coordinated interaction on large displays, in: Companion of the 2022 ACM SIGCHI Symposium on Engineering Interactive Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 26–32. URL: https://doi.org/10.1145/3531706.3536457, doi:10.1145/3531706.3536457.

Rico, J., Brewster, S., 2009. Gestures all around us: User differences in social acceptability perceptions of gesture based interfaces, in: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, Association for Computing Machinery, New York, NY, USA. pp. 1–2. URL: https://doi.org/10.1145/1613858.1613936, doi:10.1145/1613858.1613936.

Riedel, A.S., Weeks, C.S., Beatson, A.T., 2018. Am I intruding? Developing a conceptualisation of advertising intrusiveness. Journal of Marketing Management 34, 750–774. URL: https://doi.org/10.1080/0267257X.2018.1496130, doi:10.1080/0267257X.2018.1496130.

Rohrer, C., 2014. When to use which user-experience research methods. URL: https://www.nngroup.com/articles/which-ux-research-methods/.

Schrepp, M., Hinderks, A., Thomaschewski, J., 2017. Construction of a benchmark for the user experience questionnaire (UEQ). Int. J. Interact. Multim. Artif. Intell. 4, 40–44. URL: https://doi.org/10.9781/ijimai.2017.445, doi:10.9781/ijimai.2017.445.

Schrepp, M., Thomaschewski, J., 2019. Design and validation of a framework for the creation of user experience questionnaires. Int. J. Interact. Multim. Artif. Intell. 5, 88–95. URL: https://doi.org/10.9781/ijimai.2019.06.006, doi:10.9781/ijimai.2019.06.006.

Siean, A., Pamparau, C., Sluÿters, A., Vatavu, R., Vanderdonckt, J., 2023. Flexible gesture input with radars: systematic literature review and taxonomy of radar sensing integration in ambient intelligence environments. Journal of Ambient Intelligence and Humanized Computing 14, 7967–7981. URL: https://doi.org/10.1007/s12652-023-04606-9, doi:10.1007/S12652-023-04606-9.

Sluÿters, A., Sellier, Q., Vanderdonckt, J., Parthiban, V., Maes, P., 2022. Consistent, continuous, and customizable mid-air gesture interaction for browsing multimedia objects on large displays. International Journal of Human-Computer Interaction 39, 2492–2523. doi:10.1080/10447318.2022.2078464.

Valos, M.J., Maplestone, V.L., Polonsky, M.J., Ewing, M., 2017. Integrating social media within an integrated marketing communication decision-making framework. Journal of Marketing Management 33, 1522–1558. URL: https://doi.org/10.1080/0267257X.2017.1410211, doi:10.1080/0267257X.2017.1410211. number: 17-18 ZSCC: 0000044 Publisher: Routledge _eprint: https://doi.org/10.1080/0267257X.2017.1410211.

Vatavu, R.D., Wobbrock, J.O., 2015. Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1325–1334. URL: https://doi.org/10.1145/2702123.2702223, doi:10.1145/2702123.2702223.

Vermeeren, A.P.O.S., Law, E.L.C., Roto, V., Obrist, M., Hoonhout, J., Väänänen-Vainio-Mattila, K., 2010. User experience evaluation methods: Current state and development needs, in: Proceedings of the 6th Nordic Conference on Human-Computer Interaction, Association for Computing Machinery, New York, NY, USA. p. 521–530. URL: https://doi.org/10.1145/1868914.1868973, doi:10.1145/1868914.1868973.

Villarreal-Narvaez, S., Sluÿters, A., Vanderdonckt, J., Vatavu, R.D., 2024. Brave new ges world: A systematic literature review of gestures and referents in gesture elicitation studies. ACM Comput. Surv. 56. URL: https://doi.org/10.1145/3636458, doi:10.1145/3636458.

Vogiatzidakis, P., Koutsabasis, P., 2022. 'address and command': Two-handed mid-air interactions with multiple home devices. International Journal of Human-Computer Studies 159, 102755. URL: https://www.sciencedirect.com/science/article/pii/S1071581921001737,

doi:https://doi.org/10.1016/j.ijhcs.2021.102755.

Wachs, J.P., Kölsch, M., Stern, H., Edan, Y., 2011. Vision-based hand-gesture applications. Commun. ACM 54, 60–71. URL: https://doi.org/10.1145/1897816.1897838, doi:10.1145/1897816.1897838.

Walter, R., Bailly, G., Müller, J., 2013. Strikeapose: Revealing mid-air gestures on public displays, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 841–850. URL: https://doi.org/10.1145/2470654.2470774, doi:10.1145/2470654.2470774.

Wickeroth, D., Benölken, P., Lang, U., 2009. Manipulating 3d content using gestures in design review scenarios. International Journal of Information Studies 1, 242–250. URL: https://www.semanticscholar.org/paper/Manipulating-3D-Content-using-Gestures-in-Design-Wickeroth-Ben36lken, doi:https://doi.org/10.6025/ijis/2009/1/4/242-250.

Wigdor, D., Wixon, D., 2011. Brave NUI world: designing natural user interfaces for touch and gesture. Morgan Kaufmann. doi:https://doi.org/10.1016/C2009-0-64091-5. accessible at https://www.sciencedirect.com/book/9780123822314/brave-nui-world.

Williamson, J.R., 2012. User experience, performance, and social acceptability: usable multimodal mobile interaction. PhD. University of Glasgow. URL: https://eleanor.lib.gla.ac.uk/record=b2922742. zSCC: 0000014.

Wobbrock, J.O., Morris, M.R., Wilson, A.D., 2009. User-defined gestures for surface computing, in: Proc. of the ACM Int. Conf. on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 1083–1092. URL: http://doi.acm.org/10.1145/1518701.1518866, doi:10.1145/1518701.1518866.

Xia, H., Glueck, M., Annett, M., Wang, M., Wigdor, D., 2022. Iteratively Designing Gesture Vocabularies: A Survey and Analysis of Best Practices in the HCI Literature. ACM Transactions on Computer-Human Interaction 29, 1–54. URL: https://dl.acm.org/doi/10.1145/3503537, doi:10.1145/3503537.

Yee, W., 2009. Potential limitations of multi-touch gesture vocabulary: Differentiation, adoption, fatigue, in: Jacko, J.A. (Ed.), Proc. of 13th International Conference on Human-Computer Interaction, Novel Interaction Methods and Techniques, HCI' International 2009, Springer. pp. 291–300. URL: https://doi.org/10.1007/978-3-642-02577-8_32, doi:10.1007/978-3-642-02577-8_32.