LOURIM

LOUVAIN RESEARCH INSTITUTE IN MANAGEMENT AND ORGANIZATIONS

WORKING PAPER

2023/10

Definition and application of a method for evaluating the gestural interaction experience

Quentin Sellier,
Arthur Sluÿers,
Jean Vanderdonckt,
Ingrid Poncin,
Louvain Research Institute in Management and Organizations

Louvain Research Institute in Management and Organizations
Working Paper Series
Editor: Prof. Valérie Swaen
(president-lourim@uclouvain.be)

Definition and application of a method for evaluating the gestural interaction experience

Quentin Sellier, Louvain Research Institute in Management and Organizations
Arthur Sluÿters, Louvain Research Institute in Management and Organizations
Jean Vanderdonckt, Louvain Research Institute in Management and Organizations
Ingrid Poncin, Louvain Research Institute in Management and Organizations

Summary

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 modality. 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. In order to remedy this situation, this paper introduces, defines, and justifies a set of quantitative measures and three qualitative scales, i.e., discoverability, learnability, and social acceptability, to be taken into account within a global method for evaluating a gesture interface. For validation and illustration purposes, this method is applied to the gesture interface of LUI, an interactive application for browsing multimedia contents such as pictures, videos, maps, documents by a general audience. Based on this experience, we suggest some implications for the evaluation of gesture interfaces and for their incorporation in UEQ+, a modular evaluation method.

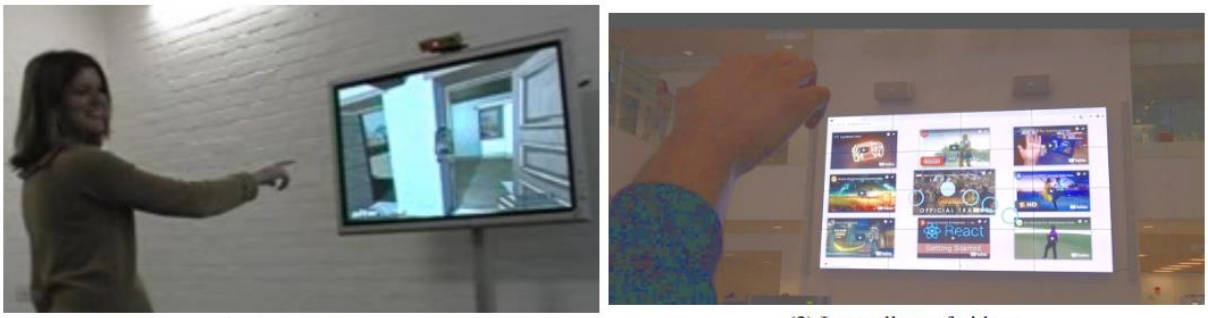
Keywords: discoverability, gesture interaction, gesture user interface, memorability, multimedia browsing, social acceptability, usability, user experience

JEL Classification: M31, O33, L86, C91

Corresponding author :
Quentin Sellier
Louvain Louvain Research Institute in Management and Organizations / Campus FUCaM Mons
Université catholique de Louvain
Chaussée de Binche 151 office B209
B-7000 Mons, BELGIUM
Email : quentin.sellier@uclouvain.be

1. Introduction

Many methods exist for evaluating the quality of a user interface (UI) in general (Dix et al, 2004 ; Vermeeren et al, 2010), but not that many specifically tailored to gesture UIs, a UI that brings into play the rich palette of all human movements for interacting with a system. For example, Figure 1 depicts a gesture UI for browsing a virtual reality 3D scene and a gallery of videos. Given the particularities of gestural interaction compared to traditional user interfaces, some authors have already put forward the need for more advanced evaluation method for gestural interaction (Wickeroth et al., 2009), capturing the multi-faced perspective of the richness of interaction (Xia et al., 2022). Rohrer (2014)'s classification distinguishes between methods producing qualitative measures capturing attitudinal UI quality attributes, such as preference, and methods producing quantitative measures expressing behavioral UI quality attributes, such as performance. This distinction is important since attitudinal measures represent end users' self-reported measure when interacting with the UI while behavioral measures report what end users really 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).



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

Figure 1: Gesture user Interfaces for multimedia browsing

Most UI evaluation methods (Vermeeren et al., 2010) target *usability*, one of the main software quality factors (ISO, 2019), that is decomposed into three sub-factors (ISO, 2018): *effectiveness* refers to task performance by expressing how accurately and completely did the user achieve the goals (e.g., measured by task completion rate and error rate), *efficiency* refers to the amount of effort that is required to achieve the level of effectiveness when achieving the goals (e.g., measured by task completion time and fatigue), *satisfaction* refers to how comfortable he user feels while using the system (e.g., assessed by a questionnaire for subjective satisfaction). There is thus an absence of subjective and affective measures, which are essential to human nature.

But currently the evaluation of user experience (UX) (Hassenzahl, 2008) goes beyond usability (van Beurden et al.,2012) because it is influenced by both the UI and its context of use (Calvary et al., 2003), which itself includes the user and the tasks, the platforms and devices, and the 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 the users' 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 just 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 modelled in a laboratory setting 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 evaluating the UX of gesture UIs. A thorough evaluation of a gesture UI should go beyond than merely assessing its usability, namely by evaluating its UX.

In principle, these evaluation methods apply to any interface, whatever their 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 the 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 and its measures within the strict framework of a gesture interface; (2) to extend an existing method to take into account new aspects not previously covered; (3) to create a new method entirely devoted to a gesture interface. 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. A combination of these options therefore seems more realistic, based on what has already been learned. This paper aims at the following contributions: to propose a method for evaluating a gesture interface by revisiting (1), extending (2), and introducing (3) qualitative and quantitative measures for such an interface. Then, to apply them on an operational interactive system with a gesture interface intended for a general public (i.e., a system for multimedia browsing on a wide variety of displays, such as large displays), to analyze these results in a detailed way and to see how these measures account for the quality of a gesture interface. In particular, we take advantage of the complementarity and convergence of approaches allowing us to capture objective and subjective measurements.

The remainder of this paper is structured as follows: Section 2 reviews the literature pertaining to UI evaluation methods with a focus on their application to gesture UIs. Section 3 states the research hypothesis of this paper and justifies why a gesture UI for multimedia browsing has been selected for conducting the evaluation. For this purpose, this section explains the design and the development of LUI, a gesture-based interactive application for multimedia browsing in multiple contexts of use. Section 4 defines and justifies our method for evaluating the user experience of a gesture UI, applies it on the LUI system, and discusses its detailed results. Section 5 synthesizes the results in terms of implications for evaluating a gesture UI by highlighting the methodological aspects of the proposed method. Section 6 concludes this paper and discusses some future avenues to this work.

2. Literature Review

In this section, we review the literature related to UI evaluation methods as soon as they have some links with gesture interaction (Section 2.1). Then, we discuss work related to gesture UI for multimedia browsing, since our LUI application belongs to this domain (Section 2.2).

2.1 Methods for Evaluating Gesture User Interfaces

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 Q1 = “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 a reasonable reliability (Cronbach’s α = .91 and .70, respectively) and they correlate with the overall SUS and with one another, they can be used separately. 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 (G1 and G4), performance (G3 and G5), and compliance with the user's expectations (G2):

- G1 = "The system recognized the gestures reliably"
- G2 = "The system followed the movements as expected"
- G3 = "The reaction time of the system is satisfactory"
- G4 = "The system aborted movements unexpectedly often"
- G5 = "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 et al. (2011) defined a quality model for evaluating a gesture UI as follows: accuracy, fatigue, and duration are objectively measured and naturalness is assessed by a 5-point Likert scale; their values are repeated for each gesture, normalized, averaged, and aggregated into a model where weights of the factors are determined by participants. This results into 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 et al. (2019) combined questionnaire-based evaluation (in this case, a 12-point rating scale for trust, a driver distraction questionnaire, the NASA TLX questionnaire (Hart & 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 on perceived performance, and the mouse scored higher on 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 et al. (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 to 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 on, Chuan et al. (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 et al. (2015) report that five evaluators discovered significantly more defects for a gesture UI with these heuristics than without. 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, few usability guidelines are suggested to evaluate a gesture UI. For instance, Yee (2009) recommends usability guidelines such as “achieve high effectiveness”, “Minimize learning among users and increase differentiation among 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 sound measure or a guideline that is empirically valid, with a strength of evidence and an importance level?

In the area of user testing, Brandao et al. (2019) evaluated a gesture UI for a tele-rehabilitation system in virtual reality 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. 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 (M = 2.23 msec) than in the gesture and vocal UI condition (M = 1.21 msec, whereas in the "wall of images" scenario, times were comparable (M = 0.58 msec vs. M = 0.61 msec).

2.2 Gesture Interaction for Multimedia Browsing

More specifically for mid-air gesture interaction with multimedia objects, less work can be found and usually focus on one media type or two at a time but not as a whole (Steinmetz & Nahrstedt, 2004): images (Neca & Duarte, 2011 ; Koutsabasis & Domouzis, 2016), videos (Zigelbaum et al., 2010), virtual objects (Ortega et al., 2017), and images with videos (Drossis et al., 2013).

For example, Ackad et al. (2015) designed and evaluated a "media ribbon", a large public display that shows information about events for theaters and concerts. Designing an intuitive and easy-to-learn gesture interface is more critical for such displays, as people are less inclined to take time to learn how to interact. They came up with four gestures (swipe left/right and move the hand up/down then hold briefly) that could trigger five different commands. To evaluate the interface, they let passers-by interact freely with the large display and interviewed them afterward. They concluded to keep gesture sets small and easy-to-learn, e.g., by always showing the most important gestures on the display and by providing context-dependent tutorials for additional gestures.

Zigelbaum et al. (2010)'s G-stalt is a chirocentric gesture interface for manipulating videos in a 3D graphical space. It featured an extensive gesture set of 21 gestures based on g-speak and relied on a Vicon motion capture system to track passive IR retro-reflective dots placed on the hands for gesture recognition, which created 3D point clouds. The actual interface consisted of a cubical arrangement of media such as photos and videos, which the user could sort through, play, and reorganize the structure into meaningful arrangement.

3. Conceptual Framework

This section defines our general research model for evaluating a gesture UI and describes LUI, a gesture UI selected to apply this model.

3.1 Definition of the Method

Each evaluation session begins with a free discovery phase, allowing to assess the discoverability of the gestures implemented. The interface must be divided into tasks each corresponding to a gesture. These are then shown to the participants for them to learn. After a break, the participant is invited, in a testing phase, to perform the requested actions. In order to test long-term memory, it is suggested to organize another session later with the same participants and the same tasks. At the end of each session a questionnaire containing the relevant scales is given to the participant.

The research model for evaluating a gestural 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 thinking time*: a positive real variable that measures in seconds the time needed by a participant to remember a gesture for a given task.
- *Gesture average thinking time*: a positive real variable that averages thinking times for all participants for a gesture task.
- *Gesture memorability rate*: a real variable between 0 and 100 that measures the percentage between the number of correctly remembered gestures and the total number of gestures produced.
- *Gesture recognition rate*: a real variable between 0 and 100 that measures the percentage between the number of correctly recognized gestures and the total number of gestures produced. 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. It is the system's "fault", the participant's "fault", respectively.

The second part of the model is composed of *gesture evaluation scale*: any scale Si, $\forall i \in \{1, \ldots, n\}$ decomposed in mi subscales $SCij$, $\forall j \in \{1, \ldots, mi\}$ 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 $SCij$ 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 $SCij$ employing a 7-point Likert scale with response categories "Strongly disagree" (= 1) to "Strongly agree" (= 7). For each Si, we have in addition:

- *Subscale Score* (SS): an ordinal integer variable that measures the score for each subscale $SCij$ of a scale Si, ranging from 1 = "Strongly disagree" to 7 = "Strongly agree".
- *Subscale Importance* (SI): an ordinal integer variable that measures the weight for each subscale $SCij$ of a scale Si, ranging from 1 = "Least important" to 7 = "Most important".
- *Scale Mean Score* (SMS): a real variable that measures the average score obtained on all subscales $SCij$ of a scale SCi for all participants, ranging from -3 = "Mostly negative" to +3 = "Mostly positive" computed as follows: SMS(SCi) = $(\sum_{j=1}^{m} SS(SCij\ ) \ / \ n) - 4$, $\forall i \in \{1, \dots, n\}$, where 4 is the scale median.
- *Scale Mean Importance* (SMI): a real variable that measures the average weight of importance of all subscales $SCij$ of a scale SCi for all participants, ranging from -3 = "Mostly negative" to +3 = "Mostly positive" computed as follows: SMI(SCi) = $(\sum_{j=1}^{m} SS(SIij\ ) \ / \ n)$ – 4, $\forall i \in \{1, \dots, n\}$, where 4 is the scale median.

For each evaluation scale, we can select established questionnaires provided that are empirically validated. For instance, each evaluation scale could be any scale defined in UEQ+ (Schrepp & Thomaschewski, 2019), a flexible and modular method for evaluating the UX of any UI.

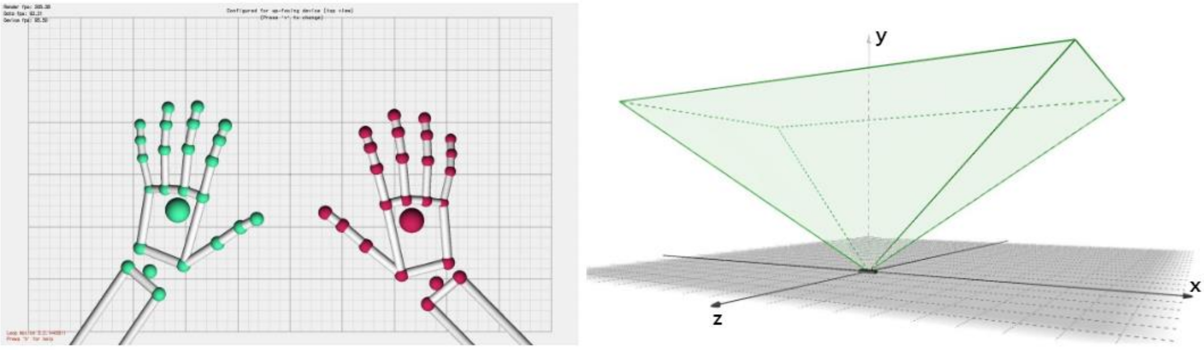
Based on these definitions, the usability could be evaluated through its three subfactors as follows: Effectiveness is covered by the *gesture task success rate* and *gesture recognition rate* variables; Efficiency is covered by the *number of gesture trials*, *gesture (average) thinking time*, *gesture memorability rate* and the *discoverability* and *learnability* scales; Subjective satisfaction is covered by the *social acceptability* scale and by any relevant *gesture evaluation scale*. Any such scale could of course supplement any of the three subfactors. UX could be covered by any *gesture evaluation scale* and its related measures, such as those factors proposed by UEQ+ (Schrepp & Thomaschewski, 2019) or the Pragmatic Quality (PQ), Hedonic Quality Stimulation (HQS), and Hedonic Quality Identification (HQI), as defined by Hassenzahl (2008).

3.2 Designing a Gesture User Interface for Multimedia Browsing

Large User Interface (LUI) is a gesture-based interactive application that enables the manipulation of multimedia objects (i.e., photos, videos, documents, and maps) on any display, such as a video wall, 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 are designed to be highly learnable and intuitive to use, as well as to impose as little constraints as possible to the user. As such, the system relies on an affordable and non-intrusive off-the-shelf device, the Leap Motion Controller (LMC), for capturing users' motion (Bachmann et al., 2018). In addition, LUI is able to automatically extract gestures from the continuous stream of motion data without additional input from the user.

The LMC plugs into a computer with USB. Using its two cameras and three infrared LEDs, it is able to accurately reconstruct a 3D skeleton of the user's hands, including the position of finger joints and hand palms within its 0.6 m wide × 0.6 m long × 0.6 m deep pyramid-shaped interaction space (cf. Figure 2). With the Leap Motion API (Ultraleap, 2022), developers can retrieve this skeleton and use it in their application, e.g., to perform hand gesture recognition (Brandon, 2014). Three versions of LUI were iteratively developed:

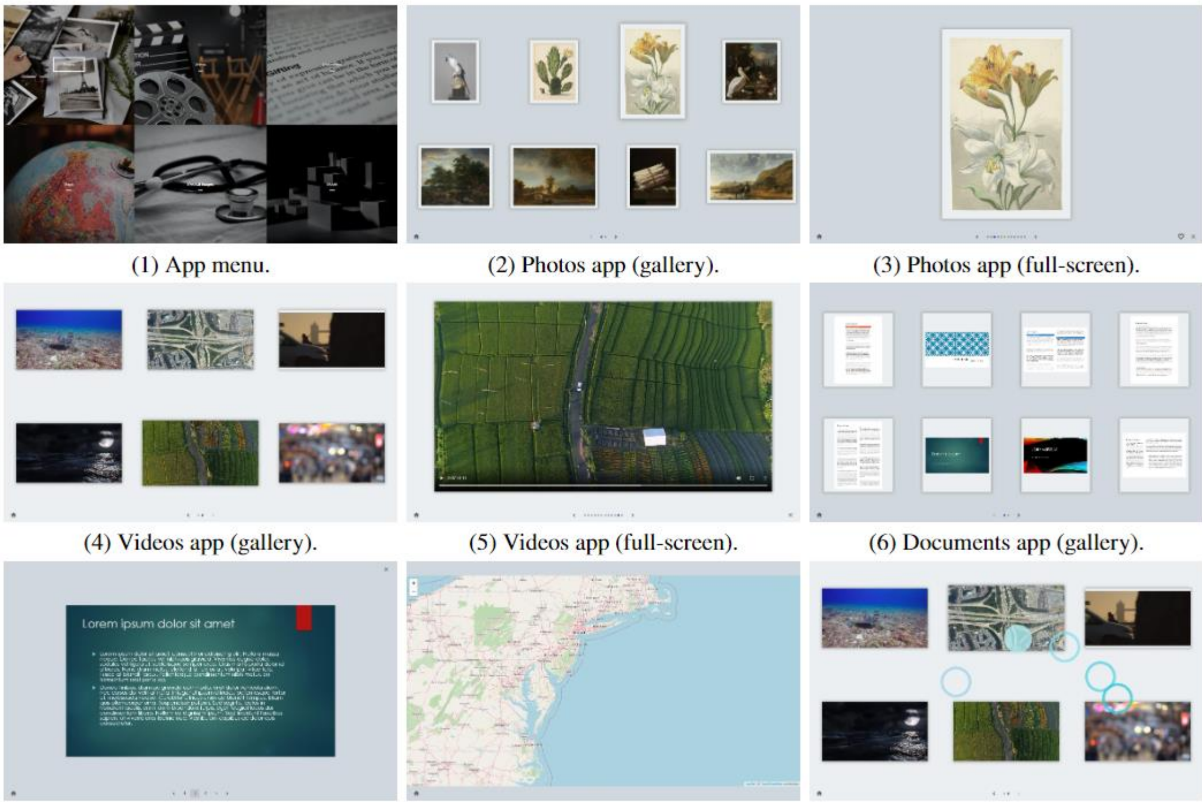
- First iteration: it featured two applications for browsing photos and videos and a limited set of gestures, such as play/pause a video, going to the next/previous item, and toggling full-screen mode.
- Second iteration: it was developed by taking into account the results of a usability testing with 20 potential end-users which highlighted some issues regarding the learning curve, accuracy of gesture recognition, and user fatigue. This version introduced simpler gestures to replace the problematic ones.
- Third iteration: it introduced many changes, including new Maps and Videos applications and gestures for manipulating content with direct feedback (e.g., zooming out of a picture by pinching the fingers). A gesture elicitation study was conducted with 23 participants to help design a highly memorable and intuitive gesture set.



(1) Hand skeleton. (2) Interaction space.

Figure 2: Illustrations of the LMC hand skeleton (1) and interaction space (2).

LUI is entirely written in JavaScript and consists of four applications (namely, Photos, Videos, Documents, and Maps) that can be accessed through an App menu (cf. Figure 3). Its implementation is divided into two main parts: a ReactJS frontend (Boduch, 2019) and a backend for gesture recognition. The backend connects to the LMC and processes the continuous flow of data from the LMC to identify user gestures. It applies filtering to the raw LMC data, performs gesture segmentation and recognition, and triggers changes in the frontend (e.g., the rotation of a picture by a couple of degrees) when relevant gestures are recognized.



(1) App menu. (2) Photos app (gallery). (3) Photos app (full-screen). (4) Videos app (gallery). (5) Videos app (full-screen). (6) Documents app (gallery). (7) Documents app (full-screen). (8) Maps app. (9) User's fingers.

Figure 3: Screenshots of the last version of the LUI application.

4. Evaluating the Gesture User Interface of LUI

An experiment was conducted following a procedure inspired by Vogiatzidakis and Koutsabasis (2022) 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 the experiment and reports its results.

4.1 Participants

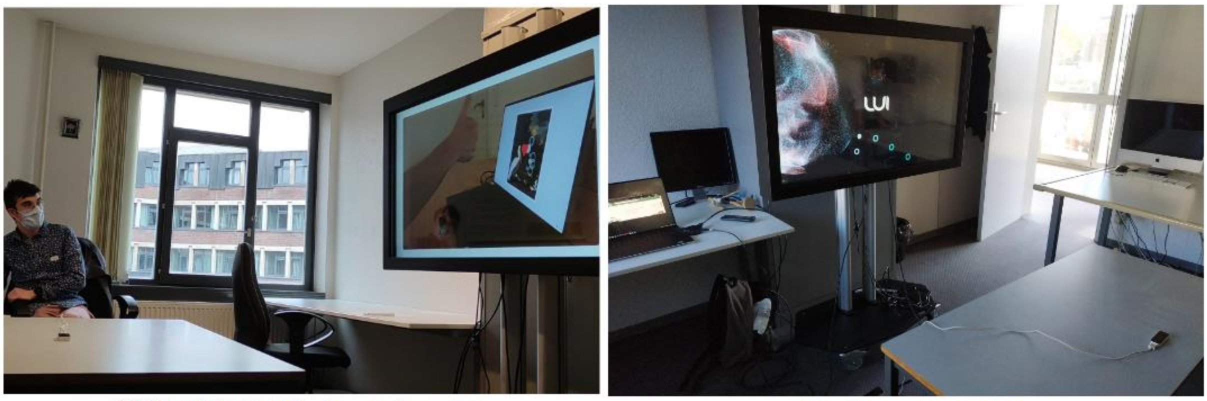
Seventeen participants (8 female, 9 male, and 0 identified as gender variant/non-conforming) aged between 22 and 65 years old (M = 37.5 years, SD = 15.3 years) were randomly selected from a list of volunteers using stratified sampling to balance between gender, background, and age (see bottom line of Table 1). 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.

Table 1: Distribution of the population sampling used in the experiment.

Id	Age (≤ or >30 y.)	Gender	Occupation (tech. *vs.* non-tech.)	Smartphone	Computer	Gesture Int.	Digital Skills
P1	34 (>)	F	Computer scientist (tech.)	7	7	1	5.75
P2	24 (≤)	M	Student (non-tech.)	7	7	1	7
P3	23 (≤)	M	Student (non-tech.)	7	7	2	6.5
P4	25 (≤)	F	PhD student (tech.)	7	7	1	5.5
P5	58 (>)	M	Accountant (non-tech.)	7	7	1	3.5
P6	22 (≤)	F	Student (non-tech.)	7	7	2	5.25
P7	55 (>)	M	Professor (tech.)	7	7	2	6.5
P8	30 (≤)	F	Communication manager (non-tech.)	7	7	1	6
P9	46 (>)	F	Computer scientist (tech.)	7	7	1	7
P10	57 (>)	F	Lab technician (non-tech.)	7	7	1	5
P11	22 (≤)	M	Student (tech.)	7	3	1	1.25
P12	23 (≤)	M	PhD student (tech.)	7	7	5	5.5
P13	36 (>)	M	Researcher (tech.)	7	7	6	6.75
P14	26 (≤)	F	PhD student (tech.)	7	7	1	3.5
P15	56 (>)	F	Civil servant (non-tech.)	7	7	1	1
P16	65 (>)	M	Physiotherapist (non-tech.)	7	4	1	2
P17	35 (>)	M	Researcher (non-tech.)	7	7	1	4.5
Mean	37.5 (8≤, 9>30 y.)	8F, 9M	8 tech., 9 non-tech	7	6.6	1.7	4.8

4.2 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 a custom software (e.g., to play a tutorial video - see Figure 4). In addition, a camera recorded both the screen and the participant's hands, and logs of the LUI application were saved on the system.



(1) Tutorial video playing on the screen.

(2) Physical configuration.

Figure 4: The setup of the experiment.

4.3 Protocol

The experiment was divided into two sessions 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, ranging from 3 to 14 days (SD = 2.2 days).

Session 1 (Discoverability and Short-term Memorability)

The objective of the first session was to introduce participants to LUI and to evaluate its discoverability, usability, as well as the short-term memorability of its gesture set. It consisted of six phases:

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, digital skills (Rose et al., 2012), and previous experience with technologies such as smartphones, computers, and gesture interaction.
2. Discovery phase. Participants interacted freely with the LUI 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 the right hand above the sensor, without touching any surface. This phase served to assess the discoverability of the gesture interface.

3. Learning phase. This phase served to teach the gestures to the participants and let them familiarize with the system. The participants were given 17 tasks in a random order (cf. Table 2). Each task was read aloud by the experimenter, who then played a short tutorial video of the associated gesture (cf. Figure 4). The use of videos ensured that all participants received the same instructions. Once the tutorial finished playing, the experimenter loaded the associated page of the LUI application and prompted the participants to perform the gesture in order to complete the task (cf. Figure 4). 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 re-loaded the correct page.
4. Resting phase. Participants were given five minutes to rest in a separate room.
5. Testing phase. This phase aimed to evaluate the short-term memorability of the gesture set by asking participants to perform 20 tasks with no external help. The participants first completed 17 atomic tasks (cf. Table 2) 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 (cf. Table 3). For each compound task, the experimenter read the task aloud, loaded the application menu, 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 the three aforementioned scales targeting the *discoverability*, *learnability*, and *social acceptability* of the LUI application (Section 4.4).

Table 2: The 17 atomic tasks, potential target content, and associated gestures.

Id	Task	Target(s)	Gestures
A1	Previous	Photo/video/document/page	Flick right
A2	Next	Photo/video/document/page	Flick left
A3	Enable fullscreen	Photo/video/document/app	Tap
A4	Disable fullscreen	Photo/video/document/app	Flick up
A5	Zoom in	Photo/document/map	Pinch out
A6	Zoom out	Photo/document/map	Pinch in
A7	Volume up	Video	Swipe up
A8	Volume down	Video	Swipe down
A9	Rotate clockwise	Photo/document	Rotate a knob clockwise
A10	Rotate anti-clockwise	Photo/document	Rotate a knob anti-clockwise
A11	Play	Video	Tap
A12	Pause	Video	Tap
A13	Like	Photo	Thumbs up
A14	Dislike	Photo	Thumbs down
A15	Fast-forward	Video	Swipe left
A16	Rewind	Video	Swipe right
A17	Pan	Map	Move hand while pointing index

Table 3: The three compound tasks and the minimum set of actions required to complete a task.

Id	Task	Instructions	Required actions
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 fullscreen, 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 fullscreen, 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 fullscreen, next, play, fast-forward

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, in particular that they could leave the experiment at any time.
2. Testing phase. This second testing phase served to evaluate the long-term memorability of the LUI gesture set. 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 & 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.

4.4 Quantitative Measures and Qualitative Scales

We collected a set of quantitative measures for the learning and testing phases of both sessions to give us insight into the user experience of the gesture UI:

- The *number of gesture trials*, as defined in our research model, measures the total number of gestures executed for each task, whether successful, incorrectly recalled, incorrectly executed or incorrectly recognized by the gesture UI.
- The *gesture task success rate*, as defined in our research model. An atomic task was considered successful if it was correctly executed within a maximum of four gesture trials. Regarding compound tasks, the number of gesture trials was not limited since they required participants to perform several gestures in an undefined order. We thus considered a compound task unsuccessful only if it was skipped by the participant.
- The *gesture thinking time*, as defined in our research model, measures the time, in seconds, needed by participants to remember a gesture for a given task.

In addition to these measures, we collected demographic information from the participants at the start of the first session. Participants were instructed to filled in at the end of each session a UEQ+ questionnaire (User Experience Questionnaire) (Schrepp et al., 2017), a modular extension of UEQ in which we selected 7 scales (i.e., *attractiveness, efficiency, perspicuity, dependability, stimulation, usefulness,* and *intuitive use*) among 20 scales to focus on evaluating the user experience of participants interacting by gestures with LUI.

In order to include these dimensions not present in UEQ+ but particularly important in the case of gesture interaction, we have created 3 ad hoc scales. These scales, created for the framework of gesture interaction, gave very satisfactory results in a previous exploratory study, both in terms of validity, with dimensionality, and internal consistency (Cronbach's α of .914 for discoverability, .726 for learnability and .941 for social acceptability). We define these scales as follows:

- *Discoverability*: the ability to make the gestures straightforward for the end user to locate, observe, and discover how they work (Wobbrock et al., 2009). It is an important aspect as it enables to quickly navigate in a UI without going through a complete training. Nacenta et al. (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 to a gesture is of course a minimal feedback. Different mechanisms, such as based on feedup and feedforward, could let the end user discover the existence of these gestures and how to operate them. The agreement score (Vatavu & Wobbrock, 2015), expressed as the common agreement between pairs of participants, can be also considered as an indicator of discoverability. It is sometimes called guessability or approachability (Xia et al., 2022). This scale is decomposed in three subscales:

  - D1 = "I could easily guess how to use the gesture interface".
  - D2 = "It was not complicated to find out how to control the gesture interface".
  - D3 = "The gestures to control the gesture interface were easy to discover".

- *Learnability*: the ease with which a new user can learn to use the interface and achieve maximal performance gestures (Dix et al., 2004). This factor is one of the most popular in the literature (Xia et al., 2022). It is linked to the intuitiveness as it is easier to learn how to use an interface based on intuitive gesture. 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". This scale is decomposed in three subscales:
  - L1 = "It was easy to learn how the gesture interface works".
  - L2 = "The mastery of the gesture interface was not complex to acquire".
  - L3 = "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. It should indeed be noted that the 3rd item refers to a progression, and is therefore no longer relevant when the individual has mastered the method of interaction perfectly.
- *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. It depends from the social (Rico & 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, and so on (Montero et al., 2010 ; Williamson, 2012). It "is determined when the motivations to use technology compete with the restrictions of social settings" (Rico

& Brewster, 2009), therefore being influenced by the context of use, i.e., user and tasks, platforms, and environments (Calvary et al., 2003). This scale is decomposed in three subscales:

- SA1 = “I would feel comfortable interacting with the gesture interface in the presence of strangers”.
- SA2 = “I would have no problem using this gesture interface in the middle of a store”.
- SA3 = “I would not feel embarrassed to use this gesture interface in the presence of others”.

Regarding the interviews, all the session were conducted by the same interviewer to ensure consistency (Riedel et al., 2018). Once the data collection phase was over, we used thematic analysis to analyze and interpret the corpus of data, via the NVivo 11 software. 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 & Clarke, 2006). The analysis was both deductive and inductive (Valos et al., 2017) by adopting a “theory-driven” approach, with themes and codes identified in prior literature and a “data-driven approach” with other codes emerging from the data (Braun & Clarke, 2006). Finally, we captured video footage of the discovery, learning, and testing phases of the experiment.

5. Results and Discussion

5.1 Gesture Task Success Rate

Table 4 gives the success rate for the atomic and compound tasks. Only two atomic tasks have a success rate below 80% in the learning phase: tasks 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). Overall, the lower success rate of the rotating (anti-)clockwise tasks can be explained by the high sensitivity of the system to the users’ hand pose. For instance, if the users’ fingers were too far apart when performing the “grab” gesture, it was not recognized by the system. To a lesser extent, this high sensitivity may have affected the recognition of other static gestures, such as in the “fast-forward”, “rewind”, and “pan” tasks.

Table 4: Success rate for the 17 atomic and 3 compound tasks.

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

Regarding compound tasks, 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 usability of the application. 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 situated on each side of the screen.

5.2 Number of Trials

Table 5 gives the average and median number of tries needed to successfully complete the atomic tasks. These metrics are presented for the learning session and the two testing sessions, allowing in particular to study the memorability of gestures.

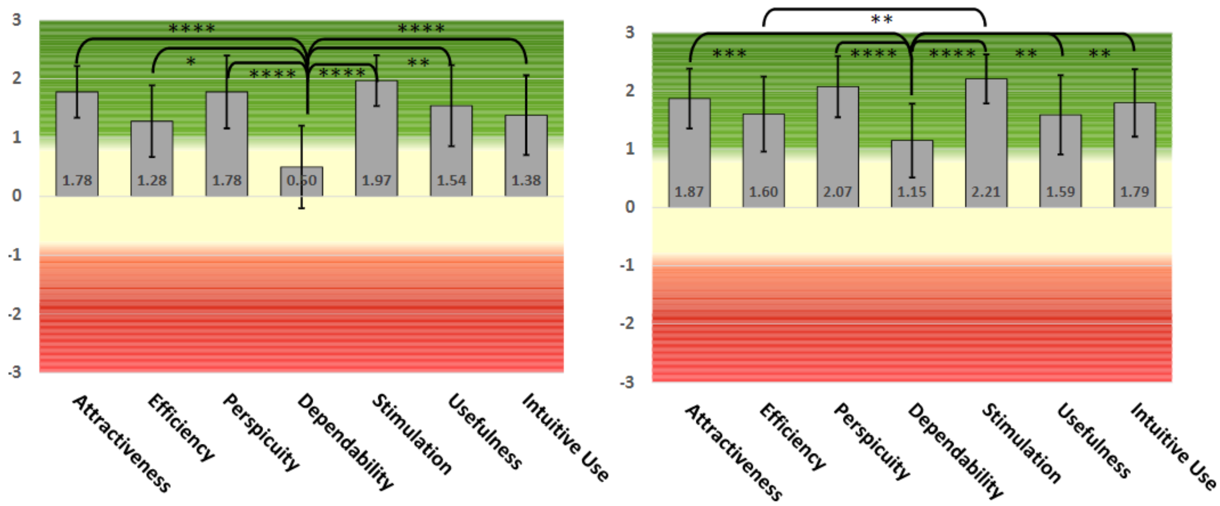
As presented, most tasks were completed with a single trial. The most problematic tasks were the rotations, whether clockwise or anticlockwise. We can also observe that more attempts were needed to pass the tasks in session 2, especially for A1 (previous), A2 (next), A12 (pause) and A15 (fast forward). On the contrary, the tasks of increasing or decreasing the volume (A7 and A8) required fewer tries over time.

Table 5: Number of trials for successful completion of each atomic task.

Id	Task	Learning phase Average	Learning phase Median	Testing phase 1 Average	Testing phase 1 Median	Testing phase 2 Average	Testing phase 2 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 fullscreen	1,5	1	1,5	1	1,1	1
A4	Disable fullscreen	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

5.3 UEQ+ Questionnaire

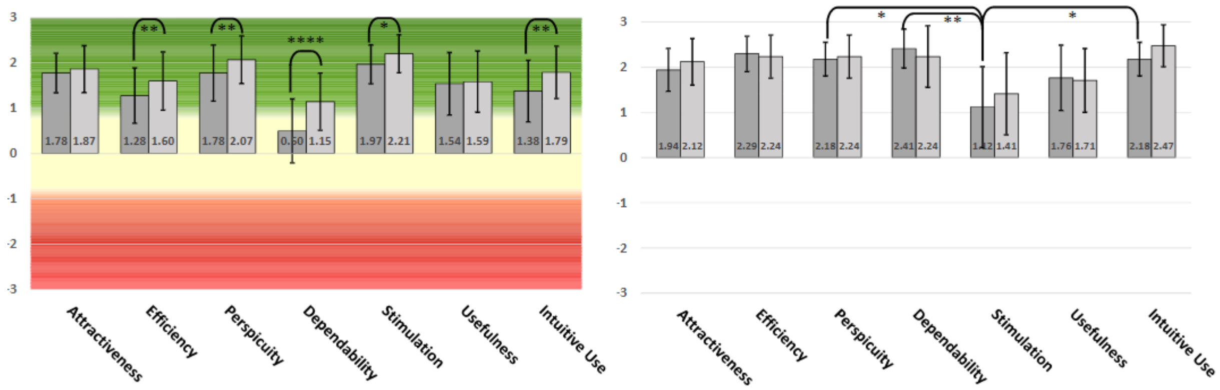
Users' answers to the seven scales selected for this study were encoded and interpreted with the UEQ data analysis tool to obtain the mean scores for the first session (cf. Figure 5–1), for the second session (cf. Figure 5–2), and for their comparison (cf. Figure 6–1). The mean score for each selected scale belongs to an interval [-3,...,+3], where a score of -3 expresses the weakest value of the scale and +3 expresses the strongest value of the scale.



(1) Session 1.

(2) Session 2.

Figure 5: Mean score for each selected scale of the UEQ+ questionnaire. Error bars show a confidence interval of 95%.



(1) Mean scores.

(2) Importance ratings.

Figure 6: Comparison of mean scores and importance ratings for each selected scale of the UEQ+ questionnaire across sessions (S1 = left in dark grey, S2 = right in light grey). Error bars show a confidence interval of 95%.

Table 6 reports the reliability of the seven scales with respect to consistency and interrating. 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 one. Interrater 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.

Table 6: Consistency reliability (Cronbach's α: G = good, V = very good, E = excellent) and Interrater reliability ((Kendall & Babington Smith, 1939)'s W : V = very weak, W = weak, M = moderate, S = strong) of mean scores (UEQ+ scales and ad hoc 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)

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 ($\alpha = 0.05$, $p < 0.0001$∗∗∗∗). 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 ($\alpha = 0.05$, $p < 0.0001$∗∗∗∗). 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. When appropriate, we also computed Spearman's rank order correlation rs(N = 17).

Intra-session Mean Scores

If we stick to the original interpretation that a mean score greater than or equal to 0.8 represents a positive evaluation (Schrepp & Thomaschewski, 2019), then all mean scores for the first session are considered positive (ranging from M = 1.28 for *efficiency* to M = 1.97 for *stimulation*), except *dependability* (M = 0.50, SD = 2.22) considered as a neutral value. This low value can be explained by the cognitive destabilization of participants producing these kinds of gestures for the first time of their life, and being puzzled by them. However, this phenomenon is observed transiently insofar as the gestures, initially felt to be difficult to produce because they are unknown, then become easier to produce. If we interpret the mean scores according to the partial benchmarking (cf. Table 1 in Schrepp et al. (2017)), then *attractiveness* is assessed as "excellent" (M = 1.78 ≥ 1.75), *efficiency* is assessed as "above average" (M=1.28 ≥ 0.98), *perspicuity* is assessed as "good" (M = 1.78 ≥ 1.56), *dependability* is assessed as "bad" (M = 0.50 < 0.78), and *stimulation* is assessed as "excellent" (M = 1.97 ≥ 1.55), as depicted in Figure 7. A Friedman test with multiple comparisons reveals that *dependability* is significantly lower than other scales during the same session (cf. Figure 5–1), for example with respect to *attractiveness* (F =146.5, p < 0.0001∗∗∗∗).

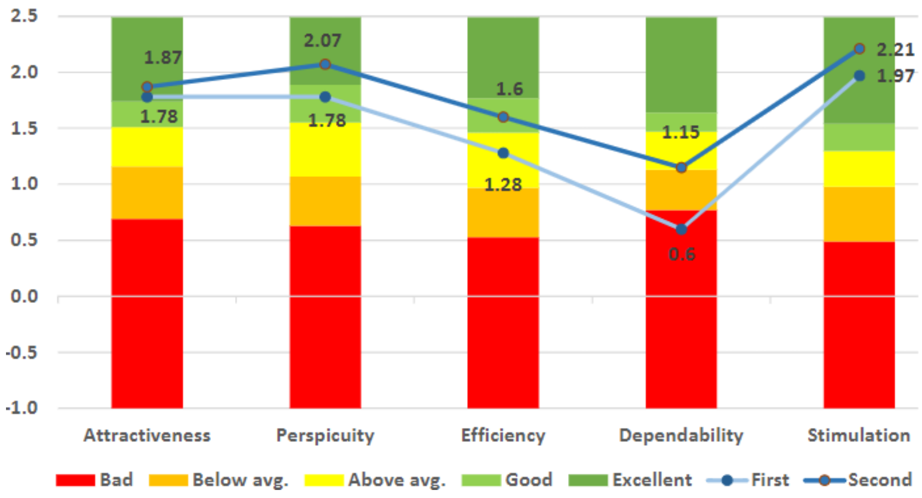


Figure 7: Benchmarking of scales for the two sessions.

Regard the second session, all mean scores are again above the threshold of 0.8 (Schrepp & Thomaschewski, 2019), ranging from a minimum for *dependability* (M = 1.15, SD = 1.33) to a maximum for simulation (M = 2.21, SD = 0.79). Mean scores resulting from this second session have all improved according to the benchmarking (Table 1 in Schrepp et al. (2017), see Figure 7): *attractiveness* is assessed as "excellent" (M = 1.87 ≥ 1.75), *efficiency* is assessed as "good" (M = 1.60 ≥ 1.60), *perspicuity* is assessed as "excellent" (M = 2.07 ≥ 1.90), *dependability* is assessed as "above average" (M = 1.15 ≥ 1.14), and *stimulation* is assessed as "excellent" (M = 2.21 ≥ 1.55). Although *dependability* increased since the first session, another Friedman test with multiple comparisons reveals that *dependability* is again significantly lower than other scales during the same session (cf. Figure 5–2), apart from *efficiency* ($F = 72$, $p = 0.089$, n.s.). However, this scale now appears to be significantly lower than *stimulation* ($F = -90$, $p = 0.0074{**}$), thereby suggesting that participants felt more stimulated than efficient by the gesture interaction.

Inter-session Mean Scores

Although *dependability* knew a negative mean score during the first session, it experienced a positive increase of 130% (M = 1.15, SD = 1.80) in the second session, which leads it to a positive evaluation. Mean scores for S2 are all superior to those obtained for S1 and range from M = 1.15 for *dependability* to M = 2.21 for *stimulation*. The mean score for some scales even largely increased, 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)).

We found a statistically significant difference for five scales among the seven used, which denotes a 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 the second session than during the first one. We also found a strong positive correlation ($rs = 0.62$, $p < 0.0001{****}$): 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.

A statistically significant difference across sessions was found for *perspicuity* across sessions (W = 263, $p = 0.0094{**}$) with a strong positive relationship ($rs = 0.69$, $p < .0001{****}$), thus suggesting that participants felt more familiar with LUI after the second session than after the first one; for *dependability* (W = 633, $p < .0001{****}$), thus suggesting that participants felt more in control of the LUI after the second session than after the first one; for stimulation (W

= 249, p = .011$*$) with a strong positive relationship (rs = 0.62, p < .0001$****$), thus suggesting that participants felt excited and motivated to use LUI in the future more after the second session than after the first one; and for *intuitive use* (W = 385, p = .011$**$) with a moderate positive relationship (rs = 0.58, p < .0001$****$), thus suggesting that participants thought that LUI could be of immediate use more after the second session than after the first one.

No significant difference was 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 on how to correctly produce gestures mainly because they never did it before. But after the second session, they felt more safe and sure about reproducing and reusing the gestures they learnt. A similar phenomenon was observed when 2D stroke gestures were introduced for the first time on Apple's iPhone: people were unsure about what gestures to do when using the interface but when they realized them, they felt more safe and able to reproduce the gestures quite efficiently.

The KPI (Key Performance Indicator) computed by UEQ+ (Hinderks et al., 2019) started from M = 1.44 (SD = 0.80) for session 1 to M = 1.74 (SD = 0.94) for session 2, which confirms the overall positive evolution.

Intra-session Importance Ratings

With regard to the first session, all importance ratings are very well consistent: the global Cronbach's α = 0.85 ≥ 0.80, all individual α are above 0.80, except *dependability* (α = 0.79 ≤ 0.80). However, the interrater reliability is very weak (Kendall & 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* (cf. Figure 7–2).

With regard to the second session, all importance ratings are very well consistent: the global Cronbach's α = 0.85 ≥ 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 Importance Ratings

No significant difference was found between the importance of scales (cf. Figure 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 first session than during the second.

5.4 Ad hoc Questionnaires

Three ad hoc scales, i.e., *discoverability, learnability*, and *social acceptability*, were added to extend the evaluation beyond the current UEQ+ scope and to cover gesture interaction. The three last rows of Table 6 show that their Cronbach's α coefficient was excellent since the beginning for *discoverability* (α = .90) until the end for *learnability* and *social acceptability* (α = .95), thus suggesting that participants responded in a consistent way that is very reliable. The coefficient of consistency reliability is approximately the same for all scales in the end of session S2. The values of Kendall's W coefficient of concordance are interpreted mostly as "weak" or "very weak", thus indicating that there was no particular agreement among participants in the way they assessed the three scales. This is pretty much justified by the three types of profiles: participants having very diverse profiles have assessed the device in their own way that is not necessarily in agreement with other participants.

Figure 8 shows how the participants' answers to the items of these three scales are distributed. Since the distribution of the answers to the 9 questions does not follow a normal distribution (all Kolmogorov-Smirnov normality tests did not pass with α = 0.05), we computed 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. Indeed, an item could receive an average answer above the median value, but not significantly. None of the three items of *discoverability* were significantly different D1 = "I could easily guess how to use the interface" (M = 3.76, SD = 1.83, Md = 4, p = .23, n.s.), D2 = "It was not complicated to find out how to control the interface" (M = 3.59, SD = 1.78, Md = 3, p = .18, n.s.), and D3 = "The gestures to control the interface were easy to discover" (M = 4.06, SD = 1.55, Md = 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 SA2 (M = 4.71, SD = 1.74, Md = 5, p = .067, n.s.). For instance, L1 (M = 5.88, SD = 0.90, Md = 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 a cognitive destabilization in the beginning. We captured this scale only once since gestures need to be discovered the first time participants are confronted to. It is likely that this scale would be rated differently during a subsequent session. Fortunately, both *learnability* and *social acceptability* were assessed very positively by participants: not only the items were almost very positively assessed but significantly better during the second session with respect to the first one.

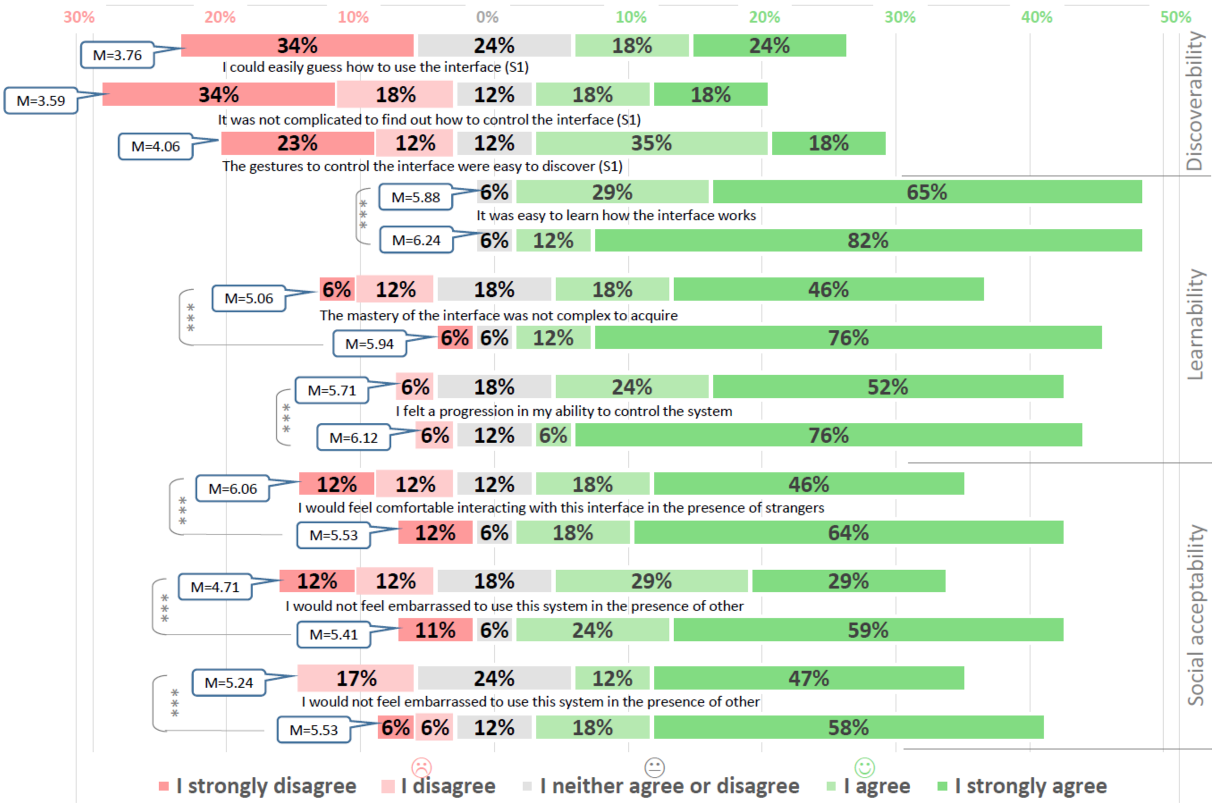


Figure 8: Distribution of participant's answers to the three new scales: discoverability (only during the first session S1), learnability (for both sessions: S1 above and S2 below), and social acceptability (for both sessions as well) (M = mean, $p < .001{*}{*}{*}$).

Indeed, 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 with respect to the first one. 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, item SA3 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-based UI during the second session that during the first one.

5.5 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. As for the advantages of this type of interface, intuitiveness was mentioned by 9 of the 17 participants in a positive way.

> "If I compare it to a computer, it's true that it's 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 providing a rich consumer experience (Holbrook & Hirschman, 1982).

Regarding the first negative point that came to the minds of the participants, the sensitivity of the system, linked to the performance of the recognizer, 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, which emerged in 10 interviews, is the direction of movement to move 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. It therefore does not mean that they consider the gesture to be correct in the way which it is currently implemented. The fact of making a movement to the right to go back in the video and to the left to go forward was chosen during the gesture elicitation study. However, now that the gesture has been implemented in a real application, participants actually feel like they are grabbing the cursor and moving it. In this logic, it would be necessary to make a movement 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 an elicitation study 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 just focused 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 to take 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. In order 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.

6. Implications for the Evaluation of Gesture Interfaces

UEQ+ (Schrepp & Thomaschewski, 2019) is an evaluation method consisting of several scales decomposed into four items each. This method, by its modularity and flexibility, represents a suitable candidate for its extension with three new scales that we define as follows: the scale *discoverability* is covered by the items "I could easily guess how to use the gesture interface", "It was not complicated to find out how to control the gesture interface", "The gestures to control the interface were easy to discover", and "I did not need help to know how to use the gesture interface"; the scale *learnability* is covered by the items "It was easy to learn how the gesture interface works", "The mastery of the gesture interface was not complex to acquire", "I felt a progression in my ability to control the gesture interface", and "Learning the gesture interface went well"; the scale *social acceptability* is covered by the items "I would feel comfortable interacting with this gesture interface in the presence of strangers", "I would have no problem using this gesture interface in the middle of a store", "I would not feel embarrassed to use this gesture interface in the presence of others", and "I would not feel judged by using this gesture interface with other people". Based on these definitions, we have updated the UEQ+ calculation tool to select these three additional scales on demand and to incorporate them into the UEQ+ method in the same way as the other scales. However, the UEQ+ calculation tool was only intended for one evaluation session, which represents only a snapshot in time. On the contrary, we need several evaluation points, especially for *learnability*. That is why we have also implemented a multi-session extension in the calculation tool as reported in this paper.

This concrete evaluation allows us to show the advantages of this method, first in combining different types of measurements. On the one hand, the task success rate remains relatively high and the median number of gesture successes is mostly 1, 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 session 2. So even though users on average had more difficulty a week later, they were more comfortable and satisfied with the interface. Another advantage of this interface is the division between atomic and compound tasks. Focusing on the gestures themselves, we could see the weakness of the rotation movements, which is 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 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, which would not have been discovered if we had been satisfied with our objective measurements and the questionnaires.

7. Conclusion and Future Work

In this article, we have detailed a new method for evaluating gestural 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 detailed 3 new scales that are particularly relevant in the context of gestural interaction, namely learnability, discoverability and social acceptability. We then applied and validated this method with LUI, a free-hand gestural interaction interface. We then discuss our results and give key takeaways.

The added 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. Secondly, 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.

This work also presents limitations and avenues for future work. The first important point is the use of our 3 ad hoc scales, which had been chosen thanks to 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. In order to increase the ecological validity of this evaluation method, it should be applied to other interfaces and contexts. This would also allow to refine the method, for instance in term of the scales used. Another future research would be to adapt this method to other contexts and interaction mediums.

References

Ackad, C., Clayphan, A., Tomitsch, M., Kay, J., 2015. An in-the-wild study of learning mid-air gestures to browse hierarchical information at a large interactive public display, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Association for Computing Machinery, New York, NY, USA. p. 1227–1238. https://doi.org/10.1145/2750858.2807532.

Archer, D., 1997. Unspoken diversity: Cultural differences in gestures. Qualitative Sociology 20, 79–105. https://doi.org/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. https://doi.org/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. https://doi.org/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 (GW ’11), Springer, Berlin, Heidelberg. pp. 36–47. https://doi.org/10.1007/978-3-642-34182-3_4.

Boduch, A., 2019. React Material-UI Cookbook. Packt Publishing. Accessible at https://www.packtpub.com/application-development/ react-material-ui-cookbook. Boyatzis, R.E., 1998. Transforming Qualitative Information: Thematic Analysis and Code Development. SAGE. ZSCC: 0017594 Google-Books- ID: _rfClWRhIKAC.

Brandon, S., 2014. Mastering Leap Motion. Packt Publishing, Birmingham.

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. pp. 189–194. URL: https://www.taylorfrancis. com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke, 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1007/978-3-319-20886-2_14.

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

Drossis, G., Grammenos, D., Birliraki, C., Stephanidis, C., 2013. MAGIC: Developing a Multimedia Gallery Supporting mid-Air Gesture- Based Interaction and Control, in: Stephanidis, C. (Ed.), Proc. of Int. Conf. on Human-Computer Interaction - Posters’ Extended Ab- stracts, Springer, Berlin, Heidelberg. pp. 303–307. https://doi.org/10.1007/978-3-642-39473-7_61.

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. https://doi.org/10.1177/0018720818824253.

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. https://doi.org/10.1016/S01664115(08)62386-9.

Hassenzahl, M., 2008. The interplay of beauty, goodness, and usability in interactive products. Hum.-Comput. Interact. 19, 319–349. https://doi.org/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. https://doi.org/10.1086/208906.

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/iso25010?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.

Koutsabasis, P., Domouzis, C.K., 2016. Mid-air browsing and selection in image collections, in: Proceedings of the ACM International Working Conference on Advanced Visual Interfaces, Association for Computing Machinery, New York, NY, USA. p. 21–27. https://doi.org/10.1145/2909132.2909248.

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.

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. https://doi.org/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. https://doi.org/10.1145/2470654.2466142.

Neca, J., Duarte, C., 2011. Evaluation of gestural interaction with and without voice commands, in: Proceedings of IADIS In- ternational Conference Interfaces and Human Computer Interaction. 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. https://doi.org/10.1145/1836216.1836228.

Ortega, F.R., Galvan, A., Tarre, K., Barreto, A.B., Rishe, N., Bernal, J., Balcazar, R., Thomas, J., 2017. Gesture elicitation for 3d travel via multi- touch and mid-air systems for procedurally generated pseudo-universe, in: Marchal, M., Teather, R.J., Thomas, B.H. (Eds.), Proceedings of IEEE Symposium on 3D User Interfaces, IEEE Computer Society. pp. 144–153. https://doi.org/10.1109/3DUI.2017.7893331.

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. https://doi.org/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.

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

Rose, S., Clark, M., Samouel, P., Hair, N., 2012. Online customer experience in e-retailing: An empirical model of antecedents and outcomes. Journal of Retailing 88, 308–322. https://doi.org/10.1016/j.jretai.2012.03.001.

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. https://doi.org/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. https://doi.org/10.9781/ijimai.2019.06.006.

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 38. https://doi.org/10.1080/10447318.2022.2078464.

Steinmetz, R., Nahrstedt, K., 2004. Multimedia Applications. Springer, Berlin, Heidelberg. pp. 197–214. https://doi.org/10.1007/978-3-662-08876-0_9.

Ultraleap, 2022. Lmc api overview. URL: https://developer-archive.leapmotion.com/documentation/csharp/devguide/Leap_ Overview.html. [Online; accessed 28-June-2022].

Valos, M.J., Maplestone, V.L., Polonsky, M.J., Ewing, M., 2017. Integrating social media within an integrated marketing com- munication decision-making framework. Journal of Marketing Management 33, 1522–1558. 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. https://doi.org/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. https://doi.org/10.1145/1868914.1868973.

Vogiatzidakis, P., Koutsabasis, P., 2022. 'address and command': Two-handed mid-air interactions with multiple home devices. International Jour- nal of Human-Computer Studies 159, 102755. 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. https://doi.org/10.1145/1897816.1897838.

Wickeroth, D., Benölken, P., Lang, U., 2009. Manipulating 3d content using gestures in design review scenar-ios. International Journal of Information Studies 1, 242–250. https://doi.org/10.6025/ijis/2009/1/4/242-250.

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.

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. http://doi.acm.org/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. https://dl.acm.org/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. https://doi.org/10.1007/978-3-642-02577-8_32.

Zigelbaum, J., Browning, A., Leithinger, D., Bau, O., Ishii, H., 2010. G-stalt: A chirocentric, spatiotemporal, and telekinetic gestural interface, in: Proceedings of the Fourth International Conference on Tangible, Embedded, and Embodied Interaction, Association for Computing Machinery, New York, NY, USA. p. 261–264. https://doi.org/10.1145/1709886.1709939.