Empirical Grounding for the Interpretations of Natural User Interface: A Case Study on Smartpen

Baraa Alabdulwahab[1] and Effie Lai-Chong Law[2]

[1]University of Leicester, LE1 7RH Leicester, UK
[2]Durham University. DH1 3LE Durham, UK
ba184@le.ac.uk;wsnv42@durham.ac.uk

Abstract. The emergence of Natural User Interface (NUI) approximately two decades ago promised to support intuitive and multimodal interactions by leveraging human sensorimotor skills such as touching, speaking, and gazing. Despite the development and introduction of commercial NUI hardware, traditional user interfaces (e.g., GUIs) continue to dominate in many sectors, prompting inquiry into the claims of 'naturalness'. To examine this phenomenon, we investigated empirically two interpretations of naturalness: innateness and intuitiveness. The study involved asking 56 participants to complete learning tasks with a smartpen system and a laptop system representing innateness and intuitiveness, respectively. A mixed-method design was implemented to collect participants' perception and performance while using both interfaces. Results indicated that, despite the smartpen system was highly learnable, the perception of naturalness was significantly linked to participants' prior experiences rather than to innate abilities. The implications of these findings are discussed.

Keywords: Natural user interface, Smartpen, Cognitive load, Innateness, Intuitiveness

1 Introduction

Natural user interface (NUI) emerged as part of the post-WIMP design trend in the early 21st century [32]. NUI aimed to answer challenges associated with the extensive use of visual elements in graphical user interfaces (GUI) to accommodate newer functionalities; such an expansion required users to learn how to operate sophisticated GUI-based computer systems [51]. In contrast, NUI gives the promise to provide more intuitive interaction, requiring minimal training and offering multimodality that increases interaction bandwidth beyond screen boundaries by using voice, touch, gestures, and other mechanisms based on biomimicry [20]. While the trend towards intuitive interaction started as early as 1980s, it was only possible to be implemented through the technological advances achieved by the end of the 20th century and the lessons learned from the development of the earlier GUI [5]. In education, for example, the natural user interaction is thought to hold a lot of potential for technology-enhanced learning as it frees learners from handling complex GUI-related instructions as a prerequisite to access learning content and provides novel ways to explore knowledge that were not possible with screen-only systems [3]. Furthermore, the ubiquitous nature of NUIs is thought to support learning analytics by feeding it with a high volume of data actively and transparently collected from learners, making thorough analysis possible [33].

In recent years, several commercial hardware devices have been introduced as NUIs. Touch screens come at the top of the list as they allow direct manipulation of visual elements on the display surface [50]. For motion, hand-held game controllers such as Nintendo Wiimote and Sony Move can track hand motion, rotation and acceleration in 3D space. The Leap Motion Controller is hand-tracking device that enables touchless (mid-air) input for computer systems [2]. Microsoft Kinect is a full-body posture and hand gestures tracker and supports spoken commands; it has been used widely in education, entertainment, health rehabilitation and training applications [47]. Voice-based interface/interaction (VUI) technologies such as Amazon Alexa and Apple Siri are deployed in an increasing number of sectors [4]. VR and AR systems such as Head-Mounted Display (HMD) are used to support reality and immersive experience [19].

Despite the high promises and potential, the advancement of the NUI technology has faced critical challenges. The rush towards applications with the lack of good foundation has translated into serious usability issues [11, 16, 29, 30, 35]. In principle, the term 'natural' has been criticised to be ambiguous due to its scope that can either mean natural to a specific group of users or natural to humans as species [37]. This basic conceptual and terminological issue was argued to have critical consequences on design decisions due to differences in targeted end users – the source where requirements for design qualities are gathered [14]. On the other hand, the perception of naturalness from the user perspective, rather than the designer perspective, was also problematic as users' expectations were demonstrated to exceed what a NUI could afford [12, 39]. Hence, identifying 'naturalness' proved to be very pivotal for designing and evaluating natural interfaces [48].

The question of naturalness has been debated intensively in numerous theoretical studies (e.g. [14, 16, 31, 37, 39, 48]). These studies followed an argumentative methodology either to negate the correctness of the term 'natural' as in [14, 37] or to seek a sensible interpretation based on previous research in the domain [16, 39, 48]. This study, in contrast, aimed to investigate naturalness empirically by engaging participants, gauging performance, and analysing experiences. Having an empirical grounding in sync with the previous theoretical discussions could be very pragmatic for developing a holistic understanding of naturalness. In the following sections, we present the unsettled disputes related to the concept of NUI and then we present our empirical contribution to address this issue.

2 Background and Related Work

2.1 Issues with NUI definition

In search for the term 'Natural User Interface', an early work by Stratton and Dunsmore [48] stated that using hyperlinks in web pages is a NUI as "[it] mimics the way humans think … in wild leaps from idea to idea" (p. 2). Since then, the term has been used more frequently, e.g., [44, 45], as an interaction style that allows humans to interact with real as well as virtual objects in a "literally direct manipulative way" [44, p.109]. In the same vein, [6] described a natural dialog interface as something that "resembles a conversation two humans might have" (p.1). Following the emergence of new input technologies, the term NUI has become common and been used to describe devices that employ touch, voice and gesture-based interactions [51].

Nevertheless, there is a lack of a universal definition for NUI in the literature. The definition in Wikipedia addresses the user's perception: "a user interface that is effectively invisible and remains invisible as the user continuously learns increasingly complex interactions". The definition from [5] focuses on the user's skills: "natural user interface is a user interface designed to reuse existing skills for interacting directly with content" (p. 2) whereas the one from [38] highlights the automation aspect: "a type of human-machine interaction based on the automatic analysis of the user's natural behaviour. These human actions are interpreted by the machine as commands that control system operations" (p. 205).

What makes the definition of NUI challenging, in comparison to command line and GUIs, is referring the interface's principle to a free-behaving user rather than to well-known and well-bounded machine artefacts. For example, in [38]'s definition cited above, the natural human behaviour is defined as "a group of activities performed by humans in everyday life to interact with their animated and unanimated environment" (p. 205). The argument here is: the interface that is invisible, reuses previous skills or relies on natural behaviour of a human cannot be defined deterministically because it is a related concept extends out of the machine domain to an active and very diverse human domain [51]. In other words, whereas a GUI can be defined as a matrix of pixels being mapped to precise coordinates on screen and specific machine commands – all finite sets, hence the abstract mathematical definition can be transferred into a hardware design in full. In contrast, a NUI is related to indeterministic and infinite sets of human behaviours, skills and perceptions that make any mathematical abstraction either ambiguous or incomprehensive [2] at pre-implementation phase.

2.2 Interpretations of naturalness in NUI

Two themes of naturalness interpretations can be observed clearly in the literature of NUI. The first theme is naturalness that emerges from the innate human abilities such as speech, touch, gestures, facial expressions, gaze, and so on [26, 42]. This theme is the most common and highlighted by systematic reviews in the domain (e.g., [13, 24, 28]. Designing a NUI under this theme can be achieved by developing a better understanding of these homo sapiens abilities [21, 43]. The other theme is naturalness that emerges from previous experiences [5, 51] and refers to learned and well-developed skills of individuals to a level that makes applying these skills happen unconsciously (i.e., without mental effort). It also entails that individuals have developed a degree of emotional tendency towards using these skills over other alternatives [51]. Designing a NUI under this theme requires methodologies such as ethnographic observation and focus group to understand target individuals [18]. While this theme relates to the previous perspective of innate abilities to some extent, it is open towards utilising more advanced interactions beyond basic human skills. Wigdor and Wixon [51] affirmed this fact by stating "NUI requires learning" (p. 12) and by stating that a keyboard is more natural for typing than a gestural interface despite the former is considered traditional while the latter considered natural.

The framework of innate and learned skills in the previous themes comes in parallel with the 'Continuum of Knowledge' suggested by [36] for intuitive interaction. The Continuum of Knowledge has four levels: innate, sensorimotor, culture and expertise. The lower two levels (innate and sensorimotor) are the most homogeneous between humans as they are inborn or develop at very young age, while the higher two levels (culture and expertise) are acquired through life experiences and can vary between groups and communities. Nauman and colleagues [36] suggested that utilising this continuum could lead to intuition as a non-conscious process of interaction. For the sake of simplicity, we call naturalness engendered by the lower levels: *innateness*, and naturalness engendered by the higher levels: *intuitiveness*. Overall, the two themes of naturalness can be referred to as: **Innateness interpretation of NUI** and **Intuitiveness interpretation of NUI.**

Nonetheless, what remains unclear is to which extent innateness vs. intuitiveness can contribute to the perception of naturalness. This question is legitimate for designing NUIs, because the intuitiveness interpretation provides more flexibility, e.g., the previous analogy of using keyboard for digital writing in [51] while innateness seems dominant in the literature of NUI. Moreover, those adopting the intuitiveness theme [5, 51] referred to touch and gestural interfaces when they discussed strategies for NUI design. Hence, it is apparent that the NUI concept in the literature is largely influenced by novel UI technologies whereas the theoretical foundation remains uncertain.

2.3 Research objectives

The aim of this study is to investigate the extent in which innateness and intuitiveness could contribute to the perception of naturalness empirically. While intuitiveness is advocated as a source of naturalness in theoretical arguments (e.g. [36, 39]), most empirical studies rely on innate abilities [13] and do not seem to support this viewpoint. Hence, this study aimed to cover this gap by collecting subjective and objective data (i.e. mixed-method approach) from users while utilising two interfacing technologies that exemplify innateness and intuitiveness themes. Data analysis can then be used as an empirical support to the meaning of naturalness.

3 Methods

3.1 Overview

The context of the empirical study was educational: a reading task followed by answering comprehension questions while using two different interfaces. The 'innate' natural user interface utilised in this study was a smart pen and paper assisted by a touchscreen, whereas the 'intuitive' user interface was a laptop (screen and keyboard interface). The aim was to probe any behavioural change or enhancement in participants' performance while using both interfaces. Further, we explicitly asked participants about their preferred interface based on their experience. Participants' vote on their preferred interface was the golden standard in deciding naturalness, while the other analysis was to justify and rationalise this selection.

3.2 Interaction assumptions and design

Interaction assumptions: A theoretically well-grounded foundation is required to justify the innateness and intuitiveness of the interfaces employed in the study. There is evidence from cognitive science that the basic use of pen and paper (i.e., scribbling) is a gesture of self-expression and communication, and can develop naturally in children as young as two [27]. Additionally, pen-based interaction is a well-known form of natural interfaces [9] that employs gestures (pen strokes) to communicate ideas. We used this foundation to derive the basic assumption that pen and paper are legitimate tools to exhibit innateness. Equally, touching and simple navigation using touchscreen is also an innate ability according to literature [5]. On the other hand, the ability to use a traditional personal computer (PC) cannot come naturally. However, for an academic cohort who use such technology for an extended period of time, using PC becomes a familiar task. Hence, the laptop system is a legitimate tool to exhibit intuitiveness for this cohort (cf. their ICT skill level was confirmed in the pre-study survey; Section 3.4).

Interaction range: Digital reading can involve a wide range of interactions. For our experiment, setting these interactions was necessary to guarantee a comparable functionality of the two interfaces. According to [15], a typical digital reading task using a computer system with keyboard and mouse as input devices covers the following interaction: scrolling, navigation through links, text search and text input. Additionally, zooming is common in smaller screens [22]. These interactions were supported by default on the laptop interface. For the smartpen system, this entailed to have a mechanism that allows an interaction with printed elements on paper, and to have a proper display modality to show output (e.g., to open URLs or show videos). The display modality also had to satisfy innateness constraints in order to keep the whole smartpen system compliant with innateness. Having this achieved in technical development (details in Sections 3.5 and 3.6), it was possible to create an interaction design for digital reading using the smartpen system and match it with the laptop system as shown in Table 1.

Table 1. A matching between digital reading interactions of laptop and smartpen systems

Laptop System	Smartpen system
Scrolling	Paper flipping
Links navigation	Tapping over printed links, output is shown on display
Text search	Tabbing over printed text, output is shown on display
Text input	Writing
Zooming	Tapping over printed elements to get a higher resolution version on display (e.g., for images and graphs)

3.3 Procedure

The learning materials for the reading tasks were two scientific articles obtained from the NASA climate blog (https://climate.nasa.gov/ask-nasa-climate/). The articles entitled "The Climate Connections of a Record Fire Year in the U.S. West" and "Sea Level 101" have the same climate theme and the same author to ensure a comparable difficulty (NB: the difficulty level was also rated by the participants as post-study feedback; Table 2 in Section 3.7).

The experiment was split into two sessions over two consecutive days with the same process. In each session, one of the two articles was provided. Each article was split into almost equal halves of similar length, and each half was presented through a different interface (i.e., a smartpen system vs a laptop system). Participants also had to use the interface to answer ten comprehension questions (9 multiple choices + 1 free text question) after reading the text.

The same procedure was used on both days, but the order of interfaces was swapped to control the order effect. Specifically, participants were asked to use the smartpen system first and then the laptop on the first day, and the laptop first followed by the smartpen system on the second day. This arrangement was made to increase accuracy and to reduce the possible bias resulted from using a new interface (the smartpen system). A session was set to maximum two hours with about one hour for each part of the article. Performance data and subjective feedback were collected during and after each task for a mixed-method evaluation (Section 3.7). The experiment was conducted on an individual basis in a quiet, reading-friendly environment.

Few days prior to the experiment, participants were contacted by email to fill a survey concerning their demographic data. The survey covered the following items: age, gender, whether English is their first language, level of education, and skill levels of using laptop and smartphone. Also, the survey included a short test to assess participants' prior knowledge of topics covered by the articles. Participants were asked to complete the survey in their own time and send it back prior to the study to minimise the time needed to spend in the lab.

3.4 Participants

The experimental study was approved by the Ethics Committee of the University of Leicester. The recruitment process targeted university students and staff, who were approached through emails and in-person invites. The sample size was 56 of which 34 were female and 22 were male; 30 of participants were non-native English speakers, mainly international students. Participants for this study were adults from different age groups: thirty-two were 18-24 years old, eleven were 25-30, seven were 31-40, three were 41-50 and two were above 50. The study did not involve any participant with special needs or learning disabilities. Participation was voluntary, and participants received a £25 Amazon gift card as a compensation for their time. All participants were confirmed to be familiar with reading using laptop web browser.

3.5 The smartpen system

Hardware: The smartpen system involved using NeoLab smartpen (model: NWP-F50) along with the specialised coded paper. Display modality was implemented by utilising Android smartphone (model: Nokia 2.4) which worked as the host system for the smartpen to display interaction output on a 6.5” touchscreen (Fig. 1).

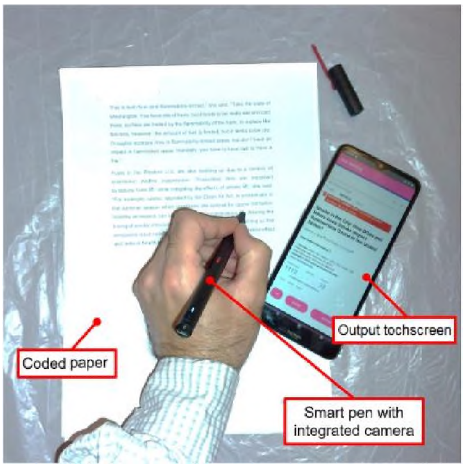

Fig. 1. The smartpen system

Software: Two software applications were developed for the smartpen system:
1) *Document procession utility:* a PC software application that takes a PDF document and converts its pages to a coded-paper PDF document so the smartpen can interact with pages upon printing. It also analyses locations of document content, and stores results in JSON format; these JSON files can be used later to interpret user input. The software can be used by authors to add/edit actions, such as playing videos or opening a URL when specific printed element is tapped.
2) *User display app:* A simple Android app for end-users to show interaction output is deployed. The app utilises the pre-generated JSON files (created by the Document processing utility) to interpret user inputs and can offer the following functionalities:
   - Displaying meaning, synonyms and translation of a (tapped) word. These data were driven directly from Google Translate website.

- Opening a URL available on the paper.
- Playing videos linked to a tapped element on paper.
- Showing a high-resolution version of images printed on paper.

3.6 The laptop system

Hardware: The laptop used in this experiment was Lenovo (model: Ideapad Yoga 13) with Windows 10 installed.

Software: A website developed to display learning materials (i.e., reading articles) followed by the comprehension questions (Fig. 2).

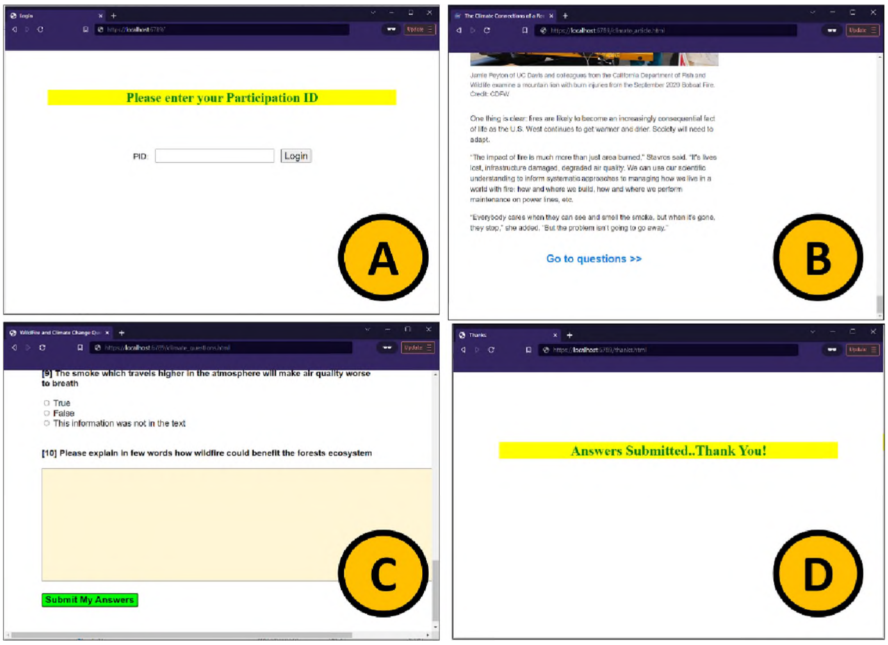

Fig. 2. Snapshots of the website designed for laptop-based reading. A: Login page, B: Reading topic page, C: Comprehension questions page, D: Submission confirmation page

3.7 Mixed-method evaluation approach

Quantitative evaluation: For quantitative evaluation, we employed the usability framework with the key metrics effectiveness and efficiency, according to the ISO 9241-110:2020 standard, along with the Cognitive Load Theory (CLT) [49] to evaluate participants' performance. These frameworks have a certain level of intersection as effectiveness and efficiency are objective measures of cognitive load [41].

Effectiveness was quantified through participant score in the comprehension test, while efficiency reflected in the time required to complete the task. Additionally, activity rate, which is the number of requests made by participant to retrieve external digital content (e.g., opening a URL or requesting the translation of a word) used to assess engagement during the reading task.

For cognitive load (CL) assessment, CLT proposes that the overall CL has three sub-components: intrinsic (ICL), extraneous (ECL), and germane (GCL) [49]. According to [10], ICL is resulted from the inherent difficulty of the learning topic, thence, it cannot be influenced by the instructional design. In contrast, ECL is resulted from the poor presentation of the learning content, while GCL is a positive mental effort that marks building knowledge in the long-term memory (i.e., creating schemas) and can be encouraged by good instructional design. Therefore, quality of learning can be indicated by a decreased ECL and increased GCL. Accordingly, a subjective assessment of the overall CL and the sub-components (ICL, ECL and GCL) was used to evaluate the quality of learning while using the two interfaces. The assessment was conducted through a questionnaire which participants had to fill after each reading task. A total of four CL questionnaires collected from each participant (2 reading tasks x 2 days). One item from [40] was used to measure the overall CL and three items from [8] were used to measure ICL, ECL and GCL, respectively. Each of the items was rated with a 9-point Likert scale (1: very very low, …, 5: neither low nor high, … 9: very very high). The items are listed as follows:

- How would you rate the <u>mental effort</u> you have invested in studying the article?
- How would you rate the <u>difficulty of the content</u> of the article?
- How would you rate the <u>difficulty to learn with the devices</u> provided?
- How would you rate <u>your concentration (attention)</u> during the reading task?

The main research question behind quantitative evaluation is whether the smartpen system can enhance the overall performance more than the laptop system. Independent variables (IVs) and dependent variables (DVs) listed in Table 2 were used, and hypotheses were formulated and verified through within-group and between-group experimental design [23] (Table 3).

Table 2. Independent and dependent variables for hypothesis testing

Independent variables	
UI Type	The type of interface being used during the task; (Smartpen, Laptop)
IT Skill level	Self-reported technical skills for using a PC and an Android device; (3 levels: Low, Medium, High)
Mother tongue	English is participant's mother language?; (Non-native, Native)
Dependent variables	
CL, ICL, ECL, GCL	Cognitive load(s) rating on a 9-point Likert scale; (Very-very low … Very-very High)

Activity	Number of requests for external materials (any sort of web resources) other than the main article during that task; (number ≥ 0). **Note:** for the smartpen system, Activity = Activity (URL requests) + Activity (word-lookup requests)
Efficiency	Time spent to complete a task; (time ≥ 0)
Effectiveness	Participant score in the comprehension test; (Grade [0..10])

Table 3. Hypotheses of the experiment

ID	Hypothesis (IV, Experimental design)	DV
H1	***When using smartpen, Activity, ECL and Efficiency are similar across participants form all IT Skill Levels groups. (IT Skill Level, Between-group)***	
H1.1	There is no significant difference in the Activity among participants with different IT skill levels.	Activity
H1.2	There is no significant difference in ECL among participants with different IT skill levels.	ECL
H1.3	There is no significant difference in the Efficiency among participants with different IT skill levels.	Efficiency
H2	***Participants' Performance and ECL are enhanced when using smartpen compared to the laptop system. (UI Type, Within-group)***	
H2.1	There is a significant difference in participants' Performance between using the smartpen and the laptop system.	Effectiveness
H2.2	There is a significant difference in participants' CL between using the smartpen and the laptop system.	CL
H3	***When using smartpen, Activity and GCL of non-native English speakers are higher than Activity and GCL of native speakers. (Mother tongue, Between-group)***	
H3.1	There is a significant difference in Activity between native and non-native English speakers.	Activity
H3.2	There is a significant difference in GCL between native and non-native English speakers.	GCL

H1 assumes that the smartpen system satisfies innateness description (i.e., no prior experience is required, and both novice and expert users can use it with the same level of proficiency [51]). H1 relies on an assumption that digitally skilled users can handle new technologies better than novice users. H2 and H3 assume that the smartpen system can enhance learning outcome. H3 relies on the fact that the smartpen system has an additional functionality (translation by tapping) that is particularly useful for non-native English speakers where learning material is provided in English.

Qualitative evaluation: Individual semi-structured interviews with participants were conducted at the end of the experiment (i.e., Day 2) where feedback on their experience of using the two interfaces was collected. Specifically, two major aspects of questions were asked: First, a comparison of reading and writing experience while utilising the laptop and smartpen systems in general and with specific reference to the perceived pros and cons of both systems. Second, whether they prefer to use the laptop or the smartpen system for future reading and writing.

Thematic analysis [7] was applied to the interview data. The analysis of qualitative data was conducted from the perspective of Disappearing Interface (DA). The DA concept [25] assumes that an interface has a physical presence as well as a conceptual presence and that invisible design should seek hiding these presences behind ubiquity and immersive interactivity, respectively [1]. According to [25], naturalness is achieved when the UI disappeared. Hence, the cues of presence were traced and measured during the qualitative analysis (Section 4.2).

4 Results

4.1 Quantitative data analysis

A total of 56 participants completed the experiment. The data were analysed using SPSS v28, and the analysis was applied to both days of the experiment. Results of Shapiro-Wilk tests indicated that dependent variables were not normally distributed ($p < 0.05$), nonparametric tests were used. First, we applied factorial analysis to study the effects of IVs and covariates and then studied the effect of individual IVs on DVs.

Quade's non-parametric factorial analysis. Demographic attributes (e.g., age, gender, education) can mediate the effect of IVs on DVs. As our data are non-normally distributed, parametric multi-factor ANOVA are inapplicable. Quade's non-parametric ANCOVA is an alternative [46], but it is less powerful and cannot show the interaction effect between two variables. It involves rank transformation of DVs. In our study, two attributes - IT skill for handling technology and status of being English native speaker for reading – are particularly relevant. We applied Quade's to evaluate the effects of these two attributes on Activity and CL (cf. Table 2). We also analysed the effects of age and gender on Effectiveness (comprehension test score) and Efficiency (task completion time). Results (Table 4a, b) show that none of these attributes have any significant effect on the DVs concerned.

Table 4. Quade's ANCOVA (a) IT Skill and Native Language on Activity and CL; (b) Age and Gender on Effectiveness and Efficiency (D1= Day1, D2= Day2; Smart =Smartpen)

	Comprehension Score				Task Time (Efficiency)			
	Smart-D1	**Smart-D2**	**PC-D1**	**PC-D2**	**Smart-D1**	**Smart-D2**	**PC-D1**	**PC-D2**
Age (Covariate, F)	0.07	0.01	0.23	2.22	2.39	1.27	0.88	0.24
Gender(Group, t)	-0.27	0.07	0.48	1.49	-1.55	-1.13	-0.94	-0.49
p (df = 54)	0.79	0.94	0.63	0.14	0.13	0.27	0.35	0.62
	Activity				**Overall Cognitive Load (CL)**			
	Smart-D1	**Smart-D2**	**PC-D1**	**PC-D2**	**Smart-D1**	**Smart-D2**	**PC-D1**	**PC-D2**
ITSkill (Covariate, F)	1.18	0.15	1.73	3.22	0.95	0.48	1.31	1.57
Native (Group, t)	1.09	0.39	1.31	1.79	-0.97	-0.69	-1.15	-1.25
p (df = 54)	0.28	0.70	0.19	0.08	0.34	0.49	0.26	0.22

Non-parametric tests with one IV. Results of Kruskal-Wallis tests, i.e., the test statistic H (degree of freedom) and *p* value, indicated no significant differences in Activity among participants from the three IT skill groups when they used the smartpen system: $H(2)_{day1} = 3.13$, $p_{day1} = .21$ and $H(2)_{day2} = .41$, $p_{day2} = .82$. The same was found for the extraneous cognitive load (ECL): $H(2)_{day1} = .12$, $p_{day1} = .94$ and $H(2)_{day2} = .59$, $p_{day2} = .75$, and for the efficiency (task completion time): $H(2)_{day1} = 3.16$, $p_{day1} = .21$ and $H(2)_{day2} = 2.56$, $p_{day2} = .28$.

As for comparing effectiveness (comprehension test score) and the cognitive load (CL) resulted from using the two systems evaluated (within-group), results of Wilcoxon signed rank tests indicated no significant differences in effectiveness (Z_{day1}= -.50, p_{day1}= .62) and (Z_{day2}= -.43, p_{day2}= .68), as well as for the cognitive load (Z_{day1}= -.39, p_{day1}= .70) and (Z_{day2}= -.15, p_{day2}= .88) between the systems. Indeed, the median comprehension test score was 8 out of 10 for both systems while the median CL for the smartpen system was 5, slightly below the median CL of the laptop system which was 6. Further, it is worth noting that results of Wilcoxon signed rank test showed a significant difference in the average Activity while using the two systems on both days (Z_{day1}= -5.55, p_{day1}< .001) and (Z_{day2}= -.5.12, p_{day2}< .001). Mean values of participants' Activity while using the smartpen system were 6.39 and 6.55 for Day1 and Day2, respectively, while the corresponding values of the laptop system were 1.77 and 1.73.

Finally, regarding the influence of mother tongue on activity and the germane cognitive load (GCL) while using the smartpen system, results of Mann-Whitney test showed no significant difference in Activity between native and non-native English speakers: $U_{day1}(N_{native}$= 26, $N_{non\text{-}native}$= 30) = 330.50, Z_{day1}= -.98, p_{day1}= .33 and U_{day2}(26, 30) = 382.50, Z_{day2}= -.12, p_{day2}= .90. Similarly, no significant difference was found for GCL on both days: U_{day1}(26, 30) = 350.50, Z_{day1}= -.67, p_{day1}= .50 and U_{day2}(26, 30) = 381, Z_{day2}= -.15, p_{day2}= .88. However, performing Mann-Whitney test on Activity related to words-lookup (i.e. finding meaning or translation) showed a significant difference between native and non-natives on day1: U(26, 30) = 244.50, Z= -2.43, p= .01, and on day2: U(26, 30) = 275.50, Z= -1.92, p= .05. The mean values of words-lookup Activity of non-natives were 4.93 and 3.37 for Day1 and Day2, respectively, almost double their corresponding values of natives, which were 2.31 and 1.88. Table 5 summarises the outcomes of hypothesis testing.

Table 5. Hypothesis testing results.

H1			H2		H3	
H1.1	H1.2	H1.3	H2.1	H2.2	H3.1	H3.2
Accept	Accept	Accept	Reject	Reject	Reject*	Reject

*: no significance in the overall activity between native and non-native speakers, however for words-lookup activity there was a significant difference.

4.2 Qualitative data analysis

A total of 56 voice recordings of post-experiment interviews were transcribed semi-automatically with an audio-to-text service (otter.ai) and then checked manually to generate final transcripts. Transcripts were analysed in four steps following the thematic analysis approach [7].

The first step was *extracting aspects of interest* in which participants expressed the 'pros' and 'cons' of two systems they had used during the evaluation. For example, feedback such as: *"I was able to put the article and questions in two tabs while I was reading... this helped in answering questions very quickly"* and *"I liked that it was possible to tap over words to get meanings immediately on phone"*.

The second step was to *group aspects of interest based on similarities*. Both positive and negative feedback belonging to the same aspect were put in the same group. For example, the ability to navigate between browser tabs and the ability of flipping pages smoothly were considered belonging to the navigability aspect.

The third step was *creating emergent, in-vivo codes* to represent data in each group. We were able to identify twelve such codes representing categories of aspects of interests (for detailed analysis the Supplementary Materials):

- **Portability:** the ability to access or to carry the system physically anywhere.
- **Capacity:** technical features and limits of a system, e.g., screen size, storage capacity or power requirement.
- **Unity:** a state whether the system is physically discrete over several parts, or it is an all-in-one device.
- **Searchability:** the possibility to find a specific piece of information, e.g., looking up a keyword.
- **Navigability:** the possibility to navigate through learning content or retrieving extra content from the web.
- **Correctability:** the possibility to undo and correct unwanted input, e.g., erasing and retyping text.
- **Distraction:** the distraction resulted from extra functionalities which is irrelevant to the learning task, e.g., responding to a notification or pop-up message.

- **Expressiveness:** the ability of the system to reveal an intention or an idea, e.g., 'free doodling' with the pen or to 'copy and paste' using the laptop.
- **Familiarity:** the level of experience in using the system.
- **Strain:** the level of stress resulted from using a system, e.g., stress resulted from the glare of screen.
- **Versatility:** the possibility of using a system for different purposes related to the learning task, e.g., to explore extra content or recording notes.
- **Engagement:** the level of attention can be achieved while using a system.

Table 6 shows the frequency (total = 236) and percentage in each of the twelve coded categories. The most frequent comments for codes are summarised in Table 7.

Table 6. Frequencies and percentage of Topics of Interests under the coding scheme

Topic of Interest category	Frequency	Percentage
Navigability	61	26%
Engagement	31	14%
Versatility	32	14%
Expressiveness	25	11%
Searchability	21	9%
Familiarity	15	6%
Capacity	13	5%
Distraction	12	5%
Strain	6	2%
Correctability	4	2%
Unity	3	1%

Table 7. Common reported features and comments under the coding scheme

Code	Notable Positives		Notable Negatives	
	Laptop Sys.	Smartpen Sys.	Laptop Sys.	Smartpen Sys.
Portability	Single device	Small and compact		Requires carrying lots of printout
Capacity	Larger screen with ability to zoom in, larger storage of text	Can be extended to larger A3 paper	Battery runout quickly	Tiny screen
Unity				Flipping between paper and phone
Searchability	Very easy to find words using Ctrl+F	Could open sources of information by a single click	Not all information types are searchable, e.g., images	Can't search words in pages
Navigability	Can open tabs side by side	Very easy way to navigate with a single click, typing URLs no longer required	Navigation between apps and websites is overwhelming	Need to flip pages frequently

Correctability	Very easy to edit and erase typed text			Need to scribble text and rewrite
Distraction			Too much distraction when browsing	Notifications comes from smartphone are distracting
Expressiveness	Can copy and paste or move items on screen	Able to doodle or sketch very quickly	Can't draw freely	Can't move text and need to rewrite it, which is time consuming
Familiarity	Using PC almost everywhere	Pen and paper are very basic and easy for writing and reading		
Strain			Screen glow causes eye burning and sore	
Versatility	Can do many things using a laptop	Able to explore videos and media directly from paper		Can't offer functionalities as much as a laptop
Engagement	Highlighting and annotating text are very helpful for reading	Reading from paper is much more engaging than a screen		

The last step of subjective analysis was to *create a meta-coding of aspects based on the concept of Disappearing Interfaces* [25]. This step involved splitting the coded categories into *physical disappearance codes* (PDC) representing aspects that are not related to the reading and writing task, and *conceptual disappearance codes* (CDC) representing interactivity that is related to the reading and writing task. Accordingly, we were able to define the two sets: PDC = {Portability, Capacity, Unity, Distraction, Familiarity, Strain}, CDC = {Searchability, Navigability, Correctability, Expressiveness, Versatility, Engagement}. The positive feedback (8) for codes in each set was considered to support disappearance while negative feedback (8) to deter disappearance.

To make a measurable 'rate of disappearance' using the thematic analysis, frequencies of positive and negative aspects under each category (i.e., laptop positive, laptop negative, smartpen positive, smartpen negative) were normalised by using the maximum value under the category as 100% and calculating the other three values accordingly. While this quantification approach ignores individual category's contribution to the disappearance of PDC/CDC sets, it highlights strengths and weaknesses of each interface. Fig. 3 shows the normalisation result.

As both PDC and CDC has 6 components each, it was also possible to calculate positive and negative scores (out of 6) of physical disappearance (PD) and conceptual disappearance (CD) for both interfaces as shown in Table 8. The accumulation (sum of positive and negative values) of scores might reflect the overall experience of disappearances (i.e., the concept of 'embodied interaction' in [39]).

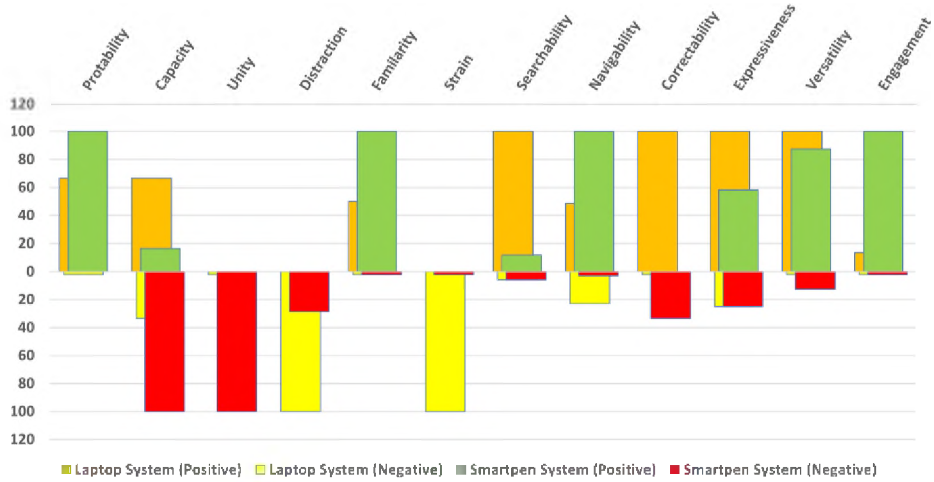

Fig. 3. Normalised frequencies of positive and negative feedback under the coding scheme

Table 8. Scores of physical and conceptual disappearances for both interfaces (pos.=positive; neg. = negative; acc. = accumulative)

	PD Score			*CD Score*		
	Pos	*neg.*	*acc.*	*pos.*	*neg.*	*acc.*
Laptop Sys.	+1.83	-2.33	***-0.5***	+4.62	-0.54	***+4.08***
Smartpen Sys.	+2.17	-2.45	***-0.28***	+3.58	-0.80	***+2.78***

With regard to participants' response on which a system they might prefer for future reading and writing, the laptop system received 34 votes for reading and 28 for writing, the smartpen system received 16 votes for reading and 25 for writing. 6 participants provided no specific preference for reading, and 3 participants provided no specific preference for writing. Voting indicated higher preference towards using a laptop for future reading and writing.

Interesting feedback received from two participants who provided no specific preference for reading as they provided a *contextual preference*. The first mentioned that she prefers paper for in-bed "relaxed" reading and the laptop for "formal" academic reading. The other participant mentioned that he prefers paper for "serious" (i.e., in-depth and highly focused) reading while the laptop for day-to-day reading. Another interesting comment from a participant who managed to navigate to the original article (available on NASA Climate website; Section 3.1) using the smartphone, and continued to read from the smartphone rather than paper of the smartpen system, she provided that "I just felt the phone *more natural* to me".

5 Discussion

The acceptance of Hypothesis 1 supported the soundness of the assumption (Section 3.2) that the smartpen system satisfies innateness description; participants demonstrated to handle the system effortlessly without previous experience. This result, along with the literature in Section 3.2, provided a firm basis for the subsequent comparisons between smartpen innateness and laptop intuitiveness.

On the other hand, the rejection of Hypothesis 2 and 3 could be interpreted in different ways. Hypothesis 2 assumed that the smartpen system would enhance performance and reduce the CL. Its rejection could be attributed either to the naturalness of the laptop system, and therefore it was able to achieve comparable effectiveness and CL levels, or to the simplicity of the task, given that all participants were academics and the median score of comprehension test was high (8/10). If we accepted the task simplicity assumption given the high scores and a modest CL level (5/9), then the naturalness of both interfaces could not be proved reliably by dependent variables in Table 2; hence, further evidence from the subjective analysis is required.

This pattern was repeated in Hypothesis 3, as the non-native participants didn't show a significant increase in the overall activity nor in GCL in comparison to their native speaking counterparts when they used the smartpen system. While the sensitivity of the CL measurement instruments adopted from [40] and [8] could influence the accuracy of CL results, the activity rate, efficiency, and effectiveness are accurate enough to confirm that no improvement in performance was associated with the use of the smartpen system. The significant increase in Activity while using the smartpen of within-group design, as well as for words lookup between native and non-natives (Section 4.1) indicated a high level of engagement. In brief, results of the quantitative analysis indicated that the smartpen system design based on innateness was able to achieve a better *interactive engagement*, but this was neither translated into a better performance nor into reduced CL. Comparable findings can be observed in the AR-JAM BOOK experiment in [17] and the experiment with Microsoft Kinect and stereoscopic visualisation of [34], and it was attributed to the difference between the *designed interaction* by developers and the *performed interaction* by participants.

The qualitative analysis of subjective feedback provided more insightful interpretations. First, the higher Familiarity of the smartpen system along with its better PD score (Fig. 3 and Table 8) support the innateness assumption. Similarly, both Engagement and Navigability support the increased activity observation. Nevertheless, the laptop system was considered to be more capable and scored much better than the smartpen system in terms of participants' perceived disappearances as indicated by the CD score. This matched the Disappearing Interface description in [25], as the physical appearance of the laptop (Distraction and Strain) was covered by the interactivity aspects (Searchability, Correctability, Expressiveness and Versatility). In other words, the laptop system which appeared physically was disappearing conceptually when participants immersed in interaction; in contrast, the lack of a comparable interactivity in the smartpen system hindered the gains in its physical disappearance.

From innateness-intuitiveness perspective, the interactivity aspects of the laptop which supported its disappearance (e.g., Searchability using CTRL+F keys; Table 7) are elements of previous *learned* experience developed in participants' perception and they don't come naturally. This supports that naturalness matches the intuitiveness description in [51] rather than innateness. A further support to this assumption can be seen by participants' recalling to the Portability aspect of the laptop vs. the smartpen system: participants were not asked during the experiment to carry and walk with the systems, but their previous experience made them associate Portability with naturalness.

The final voting and comments on system preference highlight several matters. The voting for the favour of the laptop system matched our findings of the thematic analysis, therefore it supports the soundness of the methodology. Also, the higher votes for reading using the laptop in comparison to those on writing resonates with the higher interactivity of the laptop mentioned earlier, as digital reading involves exploring different resources [15] while typical writing is more likely to be limited to a single document or writing space; the smartpen which lacked interactivity features gained comparable votes for writing. Comments from participants who provided no specific preference raises the significant aspect of *contextual naturalness*. Such feedback suggests that the exact same physical system can be natural or not natural depending on the context. This view echoes with the concept of 'embodied interaction' [39] and implies that naturalness is purely an experience in human perception rather than a feature of system. The consequences of this assumption on systems design and development could be very substantial, but it was out of our study's scope to explore them deeper.

6 Conclusion

In this study we reviewed the foundational issues of the concept of NUI, and we were able to define two possible interpretations for naturalness in the literature: innateness and intuitiveness. These interpretations were found to lack empirical support. Hence, we conducted an experimental study with the aim to compare these interpretations. Results suggested that intuitiveness is a more relevant synonym and interpretation for naturalness. Results also suggested that natural interaction could mark an experience rather than the use of a specific system. These empirical findings have substantial consequences on further NUI research. It indicates that using devices such as the Microsoft Kinect with the claim that this practice characterises an application of NUI in a specific domain (e.g., education) is no longer a legitimate approach for exploring natural interaction. Alternatively, a firm understanding to justify the selection of devices should be sought in the first place. Further, decoupling naturalness as an objective attribute of hardware and characterising it as a subjectively perceived experience implies exploring new ways for UI design, as discrete hardware modules could serve together to formulate a single NUI. It can be concluded that the NUI marks a higher level of maturity within the field of HCI rather than a specific genre of hardware. This maturity, however, was initially triggered by the emergence of novel interaction technologies.

References

[1] Arnall, T. Making Visible: Mediating the material of emerging technology, Oslo School of Architecture & Design, (2014).

[2] Bachmann, D., Weichert, F.and Rinkenauer, G. Review of three-dimensional human-computer interaction with focus on the leap motion controller. Sensors, vol. 18, no. 7, p. 2194 (2018).

[3] Becker, S., Cummins, M., Davis, A., Freeman, A., Hall Giesinger, C., and Ananthanarayanan, V. The NMC Horizon Report: 2017 Higher Education Edition. The New Media Consortium, Austin, Texas (2017).

[4] Berdasco, A., López, G., Diaz, I., Quesada, L., and Guerrero, L.A. User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana. In Proceedings of 13th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI), Toledo, Spain, (2019).

[5] Blake, J. Natural User Interfaces in .Net, Manning (MEAP) (2011).

[6] Boyce, S. J. and Gorin, A. L. User interface issues for natural spoken dialog systems. In International Symposium on Spoken Dialog (1996).

[7] Braun, V. and Clarke, V. Thematic analysis, in APA Handbook of Research Methods in Psychology, American Psychological Association, p. 57–71. (2012).

[8] Cierniak, G., Scheiter K. and Gerjets, P. Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load?. Computers in Human Behavior, vol. 25, no. 2, p. 315–324, (2009).

[9] Dai, G. Pen-based user interface, In Proceedings of The 8th International Conference on Computer Supported Cooperative Work in Design, Xiamen, (2004).

[10] Debue, N. and van de Leemput, C. What does germane load mean? An empirical contribution to the cognitive load theory," Frontier Psychology, vol. 5:1099, (2014).

[11] Farhadi-Niaki, F., Etemad, S. A. and Arya, A. Design and usability analysis of gesture-based control for common desktop tasks. In Human-Computer Interaction. Interaction Modalities and Techniques. HCI 2013. Lecture Notes in Computer Science, vol 8007, Berlin, (2013).

[12] Ghosh, S., Shruthi, C. S., Bansal, H. and Sethia, A. What is User's Perception of Naturalness? An Exploration of Natural User Experience. In Human-Computer Interaction – INTERACT, (2017).

[13] Guerino G. C. and Valentim, N. M. C. Usability and user experience evaluation of natural user interfaces: a systematic mapping study, IET Software, pp. 451-467 (2020).

[14] Hansen, L. K. and Dalsgaard, P. Note to self: stop calling interfaces "natural". In Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives, Aarhus, Denmark (2015).

[15] Hauger, D. Paramythis, A. and Weibelzahl, S. Using browser interaction data to determine page reading behavior. In Proceedings of User Modeling, Adaption and Personalization (UMAP), Girona (2011).

[16] Hornecker, E. Beyond affordance: tangibles' hybrid nature. In Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction (TEI '12), Ontario, Canada, (2012).

[17] Hornecker, E. and Dünser, A. Of pages and paddles: Children's expectations and mistaken interactions with physical–digital tools, Interacting with Computers, vol. 21, no. 1-2, pp. 95-107, (2009).

[18] Hui, T. K. L. and Sherratt, R. S. Towards disappearing user interfaces for ubiquitous computing: human enhancement from sixth sense to super senses, Journal of Ambient Intelligence and Humanized Computing, pp. 449-465 (2017).

[19] Jin, Y., Ma, M., and Zhu, Y. A comparison of natural user interface and graphical user interface for narrative in HMD-based augmented reality. Multimedia Tools and Applications, vol. 81, pp. 5795-5826, (2022).

[20] Kipper, G. and Rampolla, J. Augmented Reality: An Emerging Technologies Guide to AR, Waltham, USA: Elsevier (2012).

[21] Kortum, P. HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces, Morgan Kaufmann Publishers, (2008).

[22] Lam, H. and Baudisch, P. Summary thumbnails: readable overviews for small screen web browsers. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland Oregon, (2005).

[23] Lazar, J., Feng, J. H. and Hochheiser, H. Experimental design," in Research Methods in Human-Computer Interaction, 2nd ed., Morgan Kaufmann, , pp. 45-69, (2017).

[24] Liao, L., Liang, Y., Li, H., Ye, Y. and Wu, G. A systematic review of global research on natural user interface for smart home system, International Journal of Industrial Ergonomics, vol. 95, p. 103445, (2023).

[25] Lim, Y.-k. Disappearing interfaces, Interactions, vol. 19, no. 5, pp. 36-39, (2012).

[26] Liu, W. Natural user interface- next mainstream product user interface. In Proceedings of IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design, Yiwu, China (2010).

[27] Longobardi, C. Quaglia, R. and Iotti, N. O. Reconsidering the scribbling stage of drawing: a new perspective on toddlers' representational processes, Frontiers in Psychology, vol. 6, (2015).

[28] Loureiro, B., and Rodrigues, R. Multi-touch as a Natural User Interface for elders: A survey," in 6th Iberian Conference on Information Systems and Technologies (CISTI 2011), Chaves Portugal (2011).

[29] Luger, E. and Sellen, A. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In Proceedings of CHI Conference on Human Factors in Computing Systems, San Jose California, (2016).

[30] Macaranas, A., Antle, A. N., and Riecke, B. E. What is Intuitive Interaction? Balancing Users' Performance and Satisfaction with Natural User Interfaces. Interacting with Computers, vol. 27, no. 3, p. 357–370, (2015).

[31] Malizia, A. and Bellucci, A. The artificiality of natural user interfaces Communications of the ACM, vol. 55, no. 3, pp. 36-38, (2012).

[32] Mann, S. Natural interfaces for musical expression: physiphones and a physics-based organology," in Proceedings of the 7th International Conference on New Interfaces for Musical Expression, New York (2007).

[33] Martinez-Maldonado, R., Shum, S.B., Schneider, B., Charleer, S., Klerkx, J. and Duval, E. Learning analytics for natural user interfaces, Journal of Learning Analytics, vol. 4, no. 1, pp. 24-57 (2017).

[34] Martín-SanJosé, J.-F. Juan, M.-C. Mollá, R. and Vivó, R. Advanced displays and natural user interfaces to support learning, Interactive Learning Environments, vol. 25, no. 1, pp. 17-34, (2017).

[35] Murad, C. and Munteanu, C. I don't know what you're talking about, HALexa: the case for voice user interface guidelines. In Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin (2019).

[36] Naumann, A., Hurtienne, J., Israel, J. H., Mohs, C., Kindsmüller, M. C., Meyer, H. A. and Hußlein, S. Intuitive Use of User Interfaces: Defining a Vague Concept. In Engineering Psychology and Cognitive Ergonomics. EPCE 2007, Beijing, China (2007).

[37] Norman, D. A. Natural user interfaces are not natural. Interactions, vol. 17, no. 3, pp. 6-10, (2010).

[38] Ogiela, M. R. and Hachaj, T. Natural User Interfaces for Exploring and Modeling Medical Images and Defining Gesture Description Technology. In Natural User Interfaces in Medical Image Analysis, Springer, Cham, pp. 205-279 (2015).

[39] O'Hara, K., Harper, R., Mentis, H., Sellen, A and Taylor, A. On the naturalness of touchless: Putting the "interaction" back into NUI. ACM Transactions on Computer-Human Interaction, vol. 20, no. 1, pp. 1-25, (2013).

[40] Paas, F. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach., Journal of Educational Psychology, vol. 84, no. 4, p. 429–434, (1992).

[41] Paas, F., Tuovinen, J. E. Tabbers, H. and Van Gerven, P. W. M. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. Educational Psychologist, vol. 38, no. 1, pp. 63-71, (2003).

[42] Pavlik, J. V. Digital Design in Experiential News. In Journalism in the age of virtual reality: how experiential media are transforming news, New York, Columbia University Press, pp. 34-60, 2019,

[43] Pietroni, E., Pagano, A. and Fanini, B. UX Designer and Software Developer at the Mirror: Assessing Sensory Immersion and Emotional Involvement in Virtual Museums, Studies in Digital Heritage, pp. 13-41, (2018).

[44] Rauterberg, M. New directions in User-System Interaction: augmented reality, ubiquitous and mobile computing. In IEEE Proceedings Symposium Human Interfacing, (1999).

[45] Rauterberg, M., and Steiger, P. Pattern recognition as a key technology for the next generation of user interfaces. In IEEE International Conference on Systems, Man and Cybernetics, Beijing, China (1996).

[46] Quade, D. (1967). Rank analysis of covariance. Journal of the American Statistical Association, 62(320), 1187-1200.

[47] Sanna, A., Lamberti, F., Paravati, G., and Manuri, F. A Kinect-based natural interface for quadrotor control, Entertainment Computing, vol. 4, no. 3, pp. 179-186 (2013).
[48] Stratton, S. D. and Dunsmore, H. E. The Use of Hypertext in Software Development, SERC - Purdue University (1989).
[49] Sweller, J. Cognitive Load Theory, In Psychology of Learning and Motivation, vol. 55, Academic Press, pp. 37-76, (2011)
[50] Wigdor, D., Fletcher, J., and Morrison, G. Designing user interfaces for multi-touch and gesture devices, In Proceedings of CHI '09 Extended Abstracts on Human Factors in Computing Systems, Boston MA (2009).
[51] Wigdor, D. and Wixon, D. Brave NUI World, Morgan Kaufmann (2011).