Evaluating User Experience with physiological monitoring: A Systematic Literature Review

Article *in* Dyna (Bilbao) · March 2021

DOI: 10.6036/NT10072

3 authors:

Ainhoa Apraiz Iriarte
Mondragon Unibertsitatea

Ganix Lasa
Mondragon Unibertsitatea

Maitane Mazmela
Mondragon Unibertsitatea

EVALUATING USER EXPERIENCE WITH PHYSIOLOGICAL MONITORING: A SYSTEMATIC LITERATURE REVIEW

Ainhoa Apraiz Iriarte, Ganix Lasa Erle, Maitane Mazmela Etxabe

MONDRAGON UNIBERTSITATEA. Faculty of Engineering, Mechanical and Industrial Production, Loramendi 4, Mondragon 20500 Gipuzkoa (Spain). Tel.: +(34) 664 238 403. aapraizi@mondragon.edu

Received: 21/Jan/2021 – Reviewing: 25/Jan/2021 -- Accepted: 16/Mar/2021 - DOI: https://doi.org/10.6036/NT10072

TO CITE THIS ARTICLE:
APRAIZ-IRIARTE, Ainhoa, LASA-ERLE, Ganix, MAZMELA-ETXABE, Maitane et al. EVALUATING USER EXPERIENCE WITH PHYSIOLOGICAL MONITORING: A SYSTEMATIC LITERATURE REVIEW. DYNA New Technologies, January-December 2021, vol. 8, no. 1, p.[21 p.]. DOI: https://doi.org/10.6036/NT10072

ABSTRACT:

User Experience (UX) is a key factor and an opportunity for improvement in digital interfaces. Traditionally, it has been evaluated retrospectively through surveys and interviews. However, this is not always the optimal approach, as it does not measure UX at the moment of human-machine interaction and is therefore prone to human error due to inaccurate recall. Thus, physiological monitoring is emerging as a promising technique to assess UX during interactions. This paper aims to identify UX case studies carried out with physiological monitoring by means of a Systematic Literature Review (SLR). The results of the 33 UX case studies reviewed show that interest in incorporating physiological technologies in UX studies is growing and expanding into different fields. The electroencephalogram (EEG) was found to be the most used physiological tool, and the most used set of tools was the Galvanic Skin Response (GSR) with the electrocardiogram (ECG). In addition, the average number of participants was obtained depending on the physiological tool used. The research opportunities identified are: 1) the combination of different methods and tools in the assessment of UX, and 2), the validation of a sample size for UX tests performed with physiological monitoring.

Keywords: User Experience (UX), evaluation, Systematic Literature Review (SLR), physiological monitorization

1.- INTRODUCTION

Owing to its multidisciplinary nature, User Experience (UX) has had different meanings in its development as a field of study. In ISO 9241-210 [1] it is defined as "a person's perceptions and responses resulting from the use or anticipated use of a product, system or service". This includes user emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and achievements that occur before, during and after use [1]. Evaluation of UX is a key factor and an essential element in the quality of a product or service [2]–[4]. Evaluation refers to the application of a set of methods and instruments whose objective is to determine the perception of the use of a system or product, allowing the identification of aspects to be improved or maintained in the object of study [5].

Traditionally, UX has been evaluated retrospectively with surveys and interviews [3]. However, this is not the optimal approach in all cases, as it does not measure the experience at the time of the interaction and can be prone to human error resulting from inaccurate memory recall [6]. According to Neumann & Westbury [7] psychophysiological measures are more objective than self-report measures such as questionnaires. Furthermore, the opportunities to extract information from physiological signals are increasing due to the evolution of sensors and signal processing [8].

As stated by Maia & Furtado [9], the most commonly used psychophysiological measures in Human Computer Interaction (HCI) experiments are:

- Galvanic Skin Response (GSR)/ Electrodermal Activity (EDA), which is used to quantify an electrical potential between two points on the skin [10];
- Heart Rate (HR), which is used to measure the number of times the heart beats per minute [11];
- Electroencephalogram (EEG), which helps to assess electrical activity in the brain [12];
- Respiration [9]
- Heart Rate Variability (HRV), the variation in HR [9]
- Blood Volume Pulse (BVP), which measures changes in blood volume in arteries and capillaries [13]
- Electromyogram (EMG), which is used to measure muscle response or the response of electrical activity to stimulation of a muscle nerve [14].

Although there are many methods that use an automated approach to UX data collection [15], there are fewer examples of a well-defined method that combines automated data collection and automated UX assessment with humans. For this reason, this paper presents a systematic review of the literature on UX assessment with physiological technology tools. Furthermore, the continuous development of systems and devices, and the corresponding need for more interactive systems provides the motivation to explore the possible options for improvement [16].

This document contains the following sections: section 2 shows the related work; section 3 presents the Methodology followed to carry out the Systematic Literature Review (SLR); section 4 describes the results obtained, which are discussed in section 5; finally, section 6 presents the conclusions obtained.

2.- RELATED WORK

Maia & Furtado [15], conducted a SLR on UX evaluation. They considered that the best time for data collection is at the end of the experience, or during and after the experience. They also found that most studies had collected data manually, some in a mixed way (manual and automated) and very few automatically. The results obtained showed that measuring user emotions remains a challenge for HCI researchers [9].

Taffese [16]conducted a systematic review on UX, psychophysiology, emotion studies, the nervous system and corresponding measurement tools. The purpose of this study was to provide a solid basis for the use of psychophysiological tools in UX research. The intention was to establish a reference point, where the understanding of previous implementations and studies of psychophysiological tools, EEG and EMG in particular, is compiled. Taffese [16] extensively reviewed the advantages, limitations, contexts of use, application areas and future potentials of EEG and EMG. These tools facilitate the study of emotions in UX testing. It was concluded, however, that it is important to understand the context of use, the environment of experimentation, and the applications or software systems used when trying to extract a pattern of emotions, as the results of the tools can be ambiguous. Therefore, he recommended using a combination of psychophysiological measures and traditional measures to validate emotions.

According to the findings of Robinson, Lanius & Weber [17], UX researchers use a wide variety of methods to evaluate. They obtained similar results to Bargas-Avila & Hornbæk [18] regarding the most popular UX methods, especially surveys and interviews. The predominance of surveys, interviews, usability and focus groups suggested that UX research is largely being conducted with methods imported from other disciplines. Physiological measures have rarely been used in UX research.

The study by Baig & Kavakli [19] focused on psychophysiological signals to assess cognitive and emotional studies. Their results indicated a strong correlation between self-report data and psychophysiological data [19].

The mentioned studies are important because they present relevant results for the field about the different psychophysiological tools. To the best of the authors' knowledge, however, there is not any review which analyses how these physiological signals should be applied in user tests when assessing UX. Therefore, we decided to conduct a Systematic Literature Review (SLR), which aims to identify UX tests, usability tests or case studies conducted with physiological monitoring.

3.- METHODOLOGY

The aim of this section is to describe the process followed to carry out the Systematic Literature Review (SLR). The reason for choosing the SLR methodology is that it is considered an appropriate method to give an overview of a given field of study. Unlike conventional literature reviews, SLR follows a defined and transparent scientific process that can be replicated with the ultimate purpose of minimising reviewer bias in decisions, procedures and conclusions [20].

3.1.- SYSTEMATIC LITERATURE REVIEW PROCESS

SLR was carried out following a five-step process based on the methodology of Boell & Cecez-Kecmanovic [21] (Fig. 1). First, an SLR protocol was defined to determine the scope of the review and, in the second phase, a literature search was conducted. From the second phase, an initial database of potentially relevant literature was completed. Thirdly, the papers were reviewed and the articles that fit the objective of the review were selected. The fourth step consisted of analysing and summarising the results. Finally, the fifth step corresponds to the distribution of the data obtained.

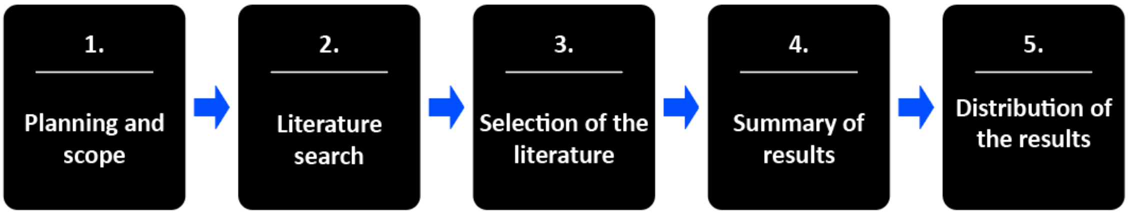

Fig. 1. Overview of the process followed to carry out the SLR. Adapted from Boell & Cecez-Kecmanovic [21].

3.2.- PLANNING THE LITERATURE SEARCH

Five databases were chosen for primary data collection. For the selection of the databases, two criteria were used [22]: 1) the databases most frequently used in the field and available in the authors' university library, and 2) the combination of different types of databases (e.g. citation databases, publisher databases and subject area databases). This variety of data sources enriches the quality of the results, as no single database covers all fields of study. Thus, data was collected from Scopus, Web of Science, Science Direct, IEEE Xplore Digital Library, and ACM Digital Library (Table 1).

Name of the Databases	Type of the Databases	Description
Scopus	Citation database	Scientific citation indexing service for citation searching of peer-reviewed journal articles. It is mainly used in Elsevier's bibliometric calculations.
Web of Science	Citation database	Scientific citation indexing service for citation searching of peer-reviewed journal articles. Mainly used in bibliometric, Thomsson/Reuter calculations.
Science Direct	Publisher database	Offers its own full-text scientific journals. Covers most disciplines, but with a focus on science, technology, and social sciences, from Elsevier and related publishers.
IEEE Xplore Digital Library	Research database	Scientific database on subjects related to computer science, electrical engineering, and electronics.
ACM Digital Library	Research database	Scientific database on subjects related to informatics and computer sciences.

Table 1: Databases selected for the SLR study.

Once the databases were selected, the next step in SLR planning was to define the literature selection criteria. All articles had to meet the following general requirements: peer-reviewed journal articles dated between January 2005 and the date this search was conducted (April 2020). Peer-reviewed journals ensure validated knowledge [23], while the year of publication limit was set to reduce the number of inappropriate hits. We did not expect to find any articles that were significant for the review before 2005.

3.2.1 – Selection of keywords

Once the databases were selected, the search equation was designed. To this end, the key words within each field were identified, as can be seen in Table 2.

Field	Keywords
Type of action	Evaluation, assessment, test, analysis
Theoretical framework	UX, "user experience", usability
Study techniques	"case study", "user test", observation, qualitative, "grounded theory"
Analysis software or tools	EEG, Eyetracking, Eyetracker, GSR, EMG, ECG

Table 2: Keywords selected for the systematic literature review.

Therefore, the search equation used in the databases was:

("User experience" OR usability OR UX) AND (evaluation OR assessment OR test OR analysis) AND ("case study" OR "user test" OR observation OR qualitative OR "grounded theory") AND (EEG OR Eyetracking OR Eyetracker OR GSR OR EMG OR ECG)

3.2.2- Location and selection of relevant studies

The execution of the literature selection process is simplified and linearly illustrated in Figure 2. Once the search terms were decided, the same filters (defined above) were applied in all databases, in order to maintain a consistent data selection process.

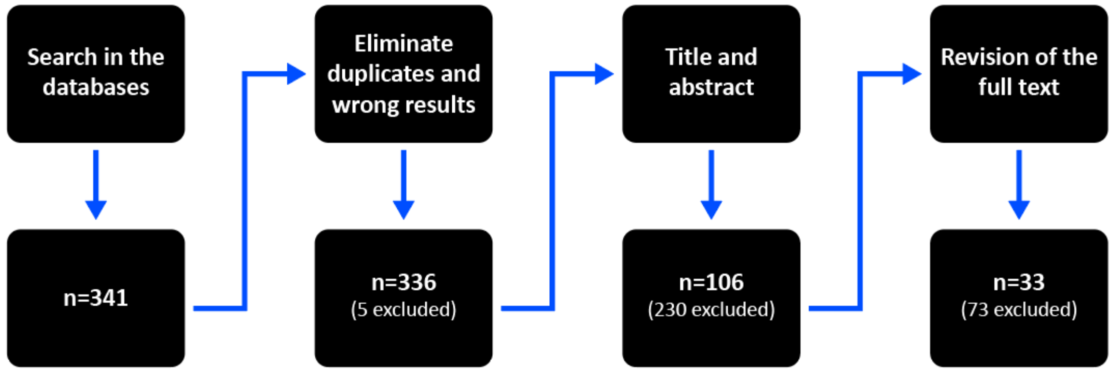

Fig. 2: Literature selection process.

The initial database search was run on the above-mentioned databases in April 2020 and a total of 341 results were obtained. The number of results per database is detailed below (Table 3).

Database	**No. of initial results**
Scopus	9
Web of Science	6
ACM	111
IEEE	137
Science Direct	78
Total	**341**

Table 3: Number of initial results per database.

Five erroneous and duplicate results were eliminated and an initial database of 336 articles was established. Erroneous articles were considered those that did not present a title or description of the papers. Articles that appeared more than once in the search results were deemed to be duplicates.

The next step consisted of a literature selection based on the article title and abstract. Inclusion/exclusion criteria were defined based on whether the article met the scope of the review (Table 4): UX tests, usability tests or case studies conducted with physiological monitoring. The scope was defined in advance to minimise uncertainty and ambiguity. However, it was common to find some articles that were unclear at this step, as sometimes the information provided by the title and abstract was insufficient to decide. In such cases, the article was accepted, as it would be re-examined in the next step with more information available. Hence, 230 documents were discarded, leaving 106 articles for later full-text review.

	Inclusion criteria
I1	Articles describing UX or usability tests or case studies using physiological monitoring.
	Exclusion criteria
E1	Articles that do NOT describe UX or usability tests or case studies using physiological monitoring.
E2	Articles in languages other than English or Spanish.
E3	Articles that are not available for download.
E4	Documents that can be considered as grey literature (e.g. course outline, thesis and dissertations), assuming that good quality grey literature research will appear in the form of journal articles [24].

Table 4: Inclusion and exclusion criteria.

All 106 full articles were reviewed, and the inclusion and exclusion criteria remained the same. At this point, a literature selection for the review was completed, finding a total of 33 journal articles within the scope.

Finally, the selected articles were analysed, and the results summarised. For this purpose, two types of data were collected: metadata and descriptive data. Metadata, such as author(s), year of publication, and source of publication, facilitated an overview of the field of interest, while descriptive data, such as scope, tools used, sample size, etc. provided a deeper qualitative understanding of the articles.

4.- RESULTS

This section summarises the results of the SLR. The section aims to provide an overview of how physiological tools are applied in UX and usability testing. It starts with a quantitative analysis and continues with a qualitative analysis. The purpose is to obtain varied and useful results for the subsequent identification of research opportunities.

4.1.- NATURE OF PUBLICATIONS

4.1.1- Evolution in the field

The first identified article dates back to 2006, when Gulliver & Ghinea [25] analysed the human side behind multimedia experiences. UX evaluation studies with physiological monitoring remained scarce over the next decade, with only 33% of articles contributing to this topic. The trend completely changed in 2016, as there have been more publications on the topic since that date. Table 5 below shows how 67% of the identified articles date from 2016 to 2020.

Year	No. of articles	% of distribution
2005	0	0%
2006	2	6%
2007	0	0%
2008	0	0%
2009	0	0%
2010	3	9%
2011	1	3%
2012	1	3%
2013	2	6%
2014	2	6%
2015	0	0%
2016	5	15%
2017	4	12%
2018	5	15%
2019	5	15%
2020	3	9%

Table 5: Number of articles identified per year.

4.1.2- Nature of the journals

As can be seen in Figure 3, the articles identified are published in 24 journals. The field of computer science is very clearly predominant (19 journals). Other journals are in the field of engineering (3 journals), one journal in the field of neuroscience, and finally one uncategorized journal (others).

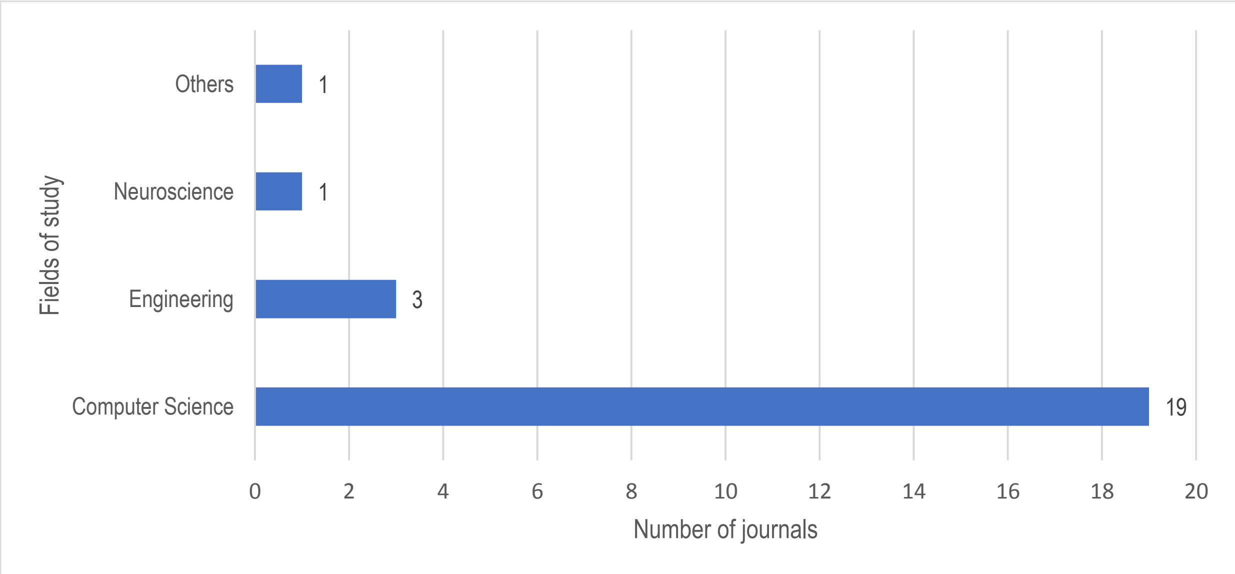

Fig. 3: Distribution of the journals in the fields of study.

The journal from which the most articles were identified is *ACM Transactions on Computer-Human Interaction* (five articles), a journal that aims to be the leading archival journal in the multidisciplinary field of human-machine interaction. Next, four articles were identified in the *IEEE Transactions on Affective Computing*. Two articles were identified in the journal *Artificial Intelligence in Medicine*, and another two in the journal *Transactions on Interactive Intelligent Systems* (Table 6).

Name of the journal	No. of articles	Journal category
ACM Transactions on Computer-Human Interaction	5	Computer Science
IEEE Transactions on Affective Computing	4	Computer Science
Artificial Intelligence in Medicine	2	Computer Science
Transactions on Interactive Intelligent Systems	2	Computer Science

Table 6: Journals in which most articles were identified.

Regarding the impact of the publications, only one article did not belong to any of the journals indexed in the Journal Citation Reports (JCR) (Table 7). In other words, 97% of the articles belonged to indexed journals. In fact, 14 journals (58%) were ranked in the first quartile (Q1), indicating the high impact of the journals.

Name of the journal	JCR Impact Factor 2019	Quartile	Category
ACM Transactions on Multimedia Computing, Communications and Applications	0.92	Q1	Computer Science
Presence	0.32	Q3	Computer Science
ACM Transactions on Accessible Computing	0.53	Q2	Computer Science
Journal of Cognitive Neuroscience	1.68	Q1	Neuroscience
Profesional de la Información	0.48	Q2	Computer Science
IEEE Transactions on Affective Computing	1.32	Q1	Computer Science
Artificial Intelligence in Medicine	1	Q1	Computer Science
Computer-Aided Design	1.05	Q1	Computer Science
WSEAS Transactions on Computers	0	Q4	Computer Science
ACM Transactions on Computer-Human Interaction	0.83	Q1	Computer Science
Personal Ubiquitous Computing	0.54	Q2	Computer Science
IEEE Transactions on Neural Systems and Rehabilitation Engineering	1.03	Q1	Computer Science
IEEE Transactions on Visualization and Computer Graphics	1.52	Q1	Computer Science
Journal of Biomedical Informatics	1.14	Q1	Computer Science
Transactions on Interactive Intelligent Systems	0.55	Q2	Computer Science
Journal of Industrial Information Integration	1.73	Q1	Engineering
Entertainment Computing	0.46	Q2	Computer Science
Studies in Health Technology and Informatics	0.27	Q3	Engineering
ACM Transactions on Sensor Networks	0.71	Q1	Computer Science
IEEE Transactions on Consumer Electronics	0.65	Q1	Engineering
Computers & Industrial Engineering	1.47	Q1	Computer Science
ACM Transactions on Privacy and Security	0.61	Q1	Computer Science
IEEE Access	0.78	Q1	Computer Science

Table 7: Impact of the journals.

4.1.3- Number of citations per article

Regarding citations of articles, the number was relatively low compared to broader fields. The most cited article was by Zickler, Halder, Kleih, Herbert, & Kübler [26] with 67 citations. In their case study EEG was used to evaluate the application, together with usability questionnaires such as NASA-TLX [27] or QUEST [28]. The work of Gulliver & Ghinea [25] was in second position (61

DYNA New Technologies, Enero-Diciembre 2021, vol. 8, no. 1, .[11 p.]. DOI: http://dx.doi.org/10.6036/NT9980

citations), and consisted of evaluating the human side of the multimedia experience by means of Eyetracking and the QoP questionnaire. With 52 citations, authors AlZoubi, D'Mello, & Calvo [29] tried to evaluate non-basic affective states using physiological technologies (EEG, EMG, and GSR) and a questionnaire where the user reports their emotions. Table 8 lists the ten most cited articles included in this review.

Ref.	Title	Author	Year	No. of citations (april 2020)
[26]	Brain painting: usability testing according to the user-centered design in end users with severe motor paralysis.	Zickler, C. Halder, S., Kleih, S. C., Herbert, C., & Kübler, A.	2013	67
[25]	Defining user perception of distributed multimedia quality.	Gulliver, S. R., & Ghinea, G.	2006	61
[29]	Detecting naturalistic expressions of nonbasic affect using physiological signals.	AlZoubi, O., D'Mello, S. K., & Calvo, R. A.	2012	52
[30]	Brain-computer interface controlled gaming: Evaluation of usability by severely motor restricted end-users.	Holz, E. M., Höne, J. Staiger-Sälzer, P., Tangermann, M., & Kübler, A.	2013	47
[31]	Automatic recognition of boredom in video games using novel biosignal moment-based features.	Giakoumis, D., Tzovaras, D., Moustakas, K., & Hassapis, G.	2011	33
[32]	Do you see what I see? The effect of gaze tracking on task space remote collaboration.	Gupta, K., Lee, G. A., & Billinghurst, M.	2016	30
[33]	A fuzzy psycho-physiological approach to enable the understanding of an engineer's affect status during CAD activities.	Liu, Y., Ritchie, J. M., Lim, T., Kosmadoudi, Z., Sivanathan, A., & Sung, R.C.	2014	20
[34]	Analysis of physiological responses to a social situation in an immersive virtual environment.	Slater, M., Guger, C., Edlinger, G., Leeb, R., Pfurtscheller, G., Antley, A., … & Friedman, D.	2006	19
[35]	A framework for physiological indicators of flow in VR games: construction and preliminary evaluation.	Bian, Y., Yang, C., Gao, F., Li, H., Zhou, S., Li, H., … & Meng, X.	2016	16
[36]	EEG-triggered dynamic difficulty adjustment for multiplayer games.	Stein, A., Yotam, Y., Puzis, R., Shani, G., & Taieb-Maimon, M.	2018	13

Table 8: Most cited articles.

4.2.- FIELD IN WHICH THE CASE STUDIES WERE CARRIED OUT

The number of articles was counted according to their field of application (Table 9). The field in which most studies were carried out is the field of app or web solutions, with 8 out of 33. This was followed by the field of Gaming, with 7 case studies. In third position were the areas of virtual reality and multimedia, which both had 3 case studies.

Field	No. of articles	Artícles
App or web solutions	8	[6], [37]–[43]
Gaming	7	[30], [31], [35], [36], [44]–[46]
Multimedia	3	[25], [47], [48]
Virtual Reality	3	[34], [49], [50]
Brain Computer Interface	2	[26], [51]
Industry 4.0	2	[52], [53]
Others	8	[29], [32], [33], [54]–[58]

Table 9: Number of articles selected by the field in which the case study was conducted.

Eight of the selected articles could not be classified in any of the fields and therefore were labelled as "other". This group included articles in fields such as machine learning [29], collaborative systems [32], *Computer-Aided Design* (CAD) [33], healthcare [54], Digital Music Instruments (DMI) [55], cognitive abilities [56], lifelogging [57], and privacy politics [58].

4.3.- MOST USED PHYSIOLOGICAL TOOL

The most frequently used physiological tool in the articles identified was EEG, which was used in 14 articles (Table 10). The second most common tool was GSR, which was employed on 11 occasions. In third position was Eyetracker, which was used in 9 articles. ECG was utilized on 8 occasions. Finally, EMG was employed on 3 occasions. In addition to these five, which were the most used, more tools were identified, however, their use was not as frequent. Studies such as those by Huynh et al. [45], Jiang et al. [57], and Peruzzini et al. [52] used wearables to monitor parameters such as photoplethysmography (PPG), HR, breathing rate (BPM) or skin temperature (ST). Table 10 shows the most frequently used tools and the articles in which each of them was included.

Physiological tool	No. of articles	Articles
EEG	14	[6], [26], [30], [33], [36], [39], [41]–[44], [47], [49], [50], [55], [57]
GSR/EDA	11	[29], [31], [33], [34], [38]–[41], [45], [46], [48], [51], [56]
Eyetracker	9	[6], [25], [32], [37], [41], [47], [52], [53], [56], [58]
ECG	8	[29], [31], [34], [35], [39], [46], [55], [56]
EMG	3	[29], [33], [35]

Table 10: Number of articles per physiological tool used.

4.4.- PHYSIOLOGICAL TOOLS MOST USED TOGETHER

In order to identify which tools were used together in the UX and usability tests, the number of times they were used together was counted (Table 11). At the top of the list, was ECG with GSR, which were used together 7 times. This was followed by ECG and EEG (3 times).

Physiological toolkit	No. of times used
ECG and GSR	7
ECG and EEG	3
Eyetracker and EEG	2
GSR and EEG	2
EMG and GSR	2
ECG and EMG	2

Table 11: Most used physiological toolkit.

4.5.- MOST USED PHYSIOLOGICAL TOOL ACCORDING THE FIELD OF APPLICATION

Figure 4 shows that the application field with the most case studies was the field of app or web solutions. In this field, the most used tool was EEG (4 times), followed by GSR/EDA, Eyetracker and ECG, which were used 3 times each. No study carried out with EMG was identified in this field. In the field of Gaming, EEG, GSR/EDA, and ECG were used 3 times. In this field, EMG was also used once. In Multimedia the most employed device was Eyetracker (2 times), and, to a lesser extent, EEG, GSR/EDA, and ECG. However, EMG was not used. EEG was the most used in virtual reality (2 times), followed by ECG and GSR/EDA. EEG was also the most utilized in Brain Computer Interface, and GSR/EDA was also used. In Industry 4.0, Eyetracker was used twice.

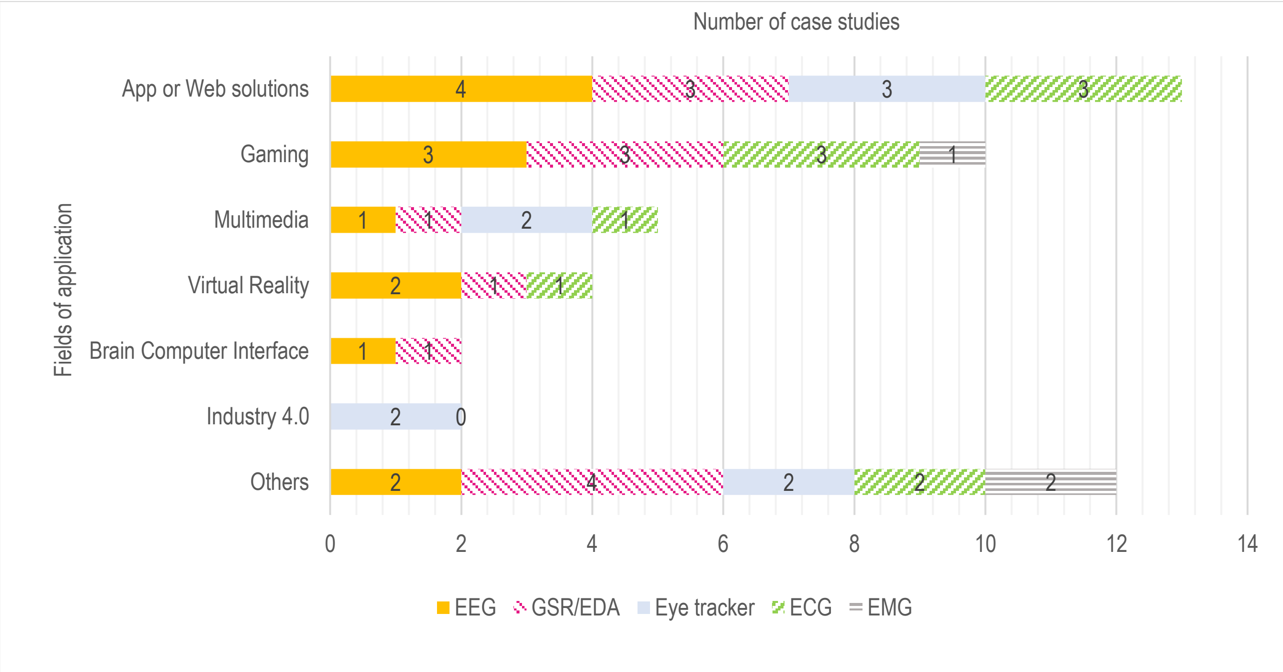

Fig. 4: Most used physiological tool according to field.

DYNA New Technologies, Enero-Diciembre 2021, vol. 8, no. 1, .[11 p.]. DOI: http://dx.doi.org/10.6036/NT9980

4.6.- MOST USED QUESTIONNAIRES

In the evaluation of user perception, the most frequently used technique was the semi-structured interview, which was repeated on 8 occasions. The most used questionnaires were SUS and NASA-TLX. However, they were not used significantly, as they were only used on 3 and 2 occasions respectively. The socio-demographic questionnaires were disregarded, as they are used to segment groups of users and not to evaluate the experience.

4.7.- SAMPLE SIZE ACCORDING TO PHYSIOLOGICAL TOOL

The mean of the sample used in the case studies was 24, and the standard deviation 22.6. This means that the sample size varied from case study to case study. The smallest sample size was 4 by Holz, Staiger-Sälzer, Tangermann, & Kübler [30], Zickler et al. [26], and Peruzzini et al. [52]. In contrast, the largest sample used was 108 by Gulliver & Ghinea [25].

The average number of participants in the case studies depending on the physiological tool used was calculated. Figure 6 summarizes the results:

- In the case of EEG, the median sample size was 18 participants and the mean value was 19.94. 50% of the EEG case studies had a sample size of 9.5 to 24.5 participants.
- In the case of GSR/EDA, the median sample size was 25 and the mean value was 26.31. 50% of the GSR/EDA case studies had a sample size ranging from 18.5 to 32.5.
- For Eyetracker, the median sample size was 18 and the mean value was 32.09. 50% of the Eyetracker case studies had a sample size ranging from 12 to 54.
- The median sample for the ECG was 25 and the mean value was 28.89. 50% of the case studies performed with ECG had a sample size of 19.5 to 35.5.
- In the case of EMG, as only 3 case studies were identified, the median was the case study with a sample size of 27, the lower limit was the one with 15, and the upper limit was the one with 36.

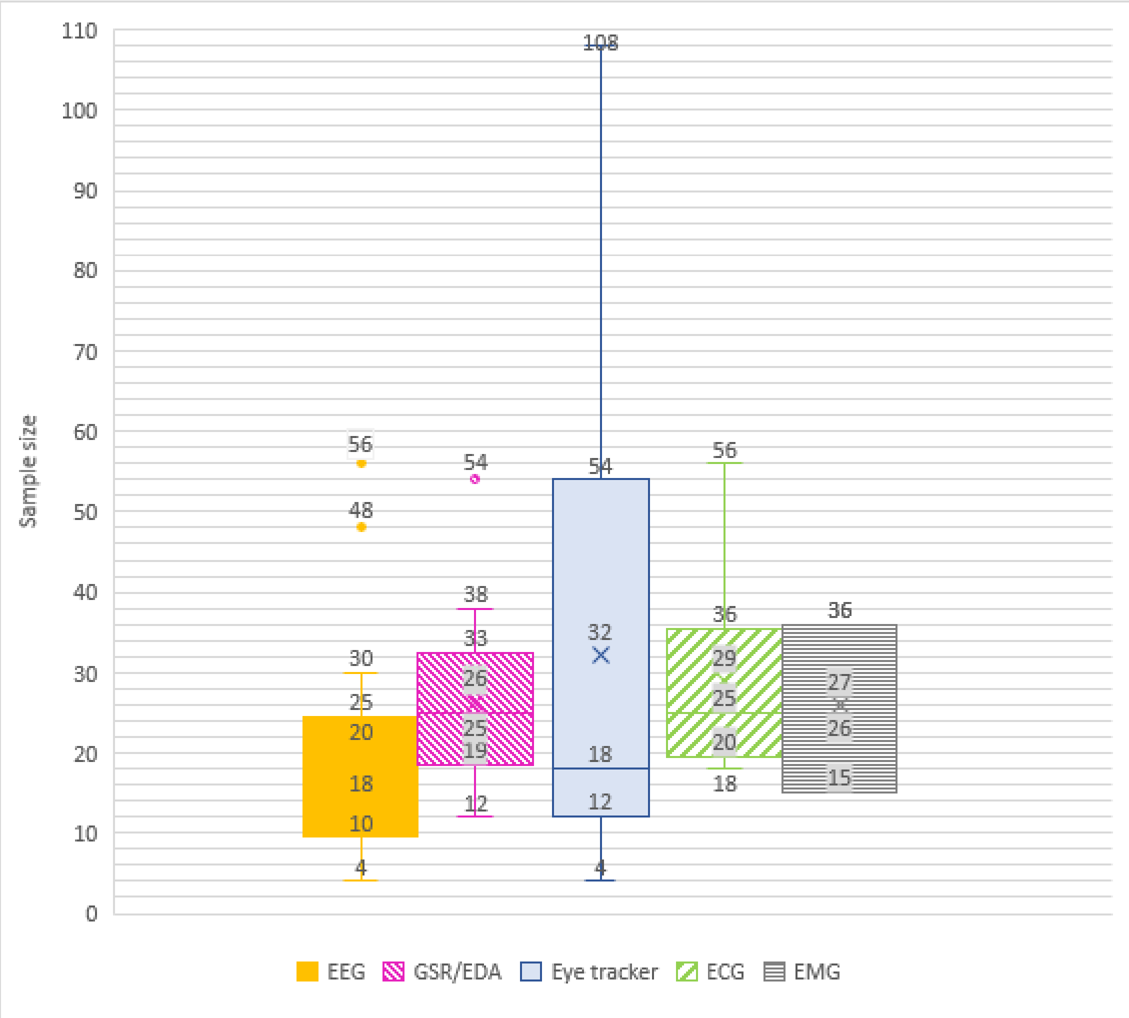

Fig. 5: Sample size according to the physiological tool.

5.- DISCUSSION AND RESEARCH OPPORTUNITIES

5.1.- EVOLVING TECHNOLOGIES

The limited number of scientific publications related to UX evaluation with physiological monitorization indicates that this is an emerging field of study. The results showed that interest in incorporating physiological technologies in UX studies is growing however, with 67% of articles identified dating from 2016 to 2020. This increase could be due to the development of the UX concept that includes different ways of evaluating the before, during and post-interaction. Another phenomenon accompanying this increase might be the fact that there has been a considerable decrease in the cost of biometric recognition systems. Implementations of such systems has seen a marked increase in recent years, driving a need for mass production and thus reducing manufacturing costs [59]. This price reduction could also be related to the increased use of this type of technology in UX testing over the last 5 years.

However, there are certain variables that could limit this increase, not least of which could be cost [16]. Although prices have decreased as a result of mass production, the cost of these systems can still be out of the reach of UX professionals. A second reason could be the time needed for setup, calibration and subsequent data analysis [60]. In addition, a high degree of expertise is needed to transform the obtained data into useful information [16]. Thirdly, the data obtained may not always be reliable, since the physical movements of the participants can cause unrealistic spikes [16]. One final reason could be the degree of intrusiveness for the participants.

These limitations, in addition to the degree of confidentiality of the sector, might account for the limited number of studies identified in the field of Industry 4.0. Furthermore, in UX or usability tests, it is common for the participant to make physical movements during the interaction. Such physical movements can cause unrealistic data spikes however, making subsequent analysis difficult. In fact, EEG has not been used in any of the Industry 4.0 case studies, since, in addition to making data recording and analysis more difficult, it can be considered more disruptive and limit participants' movements. The GSR, on the other hand, is a device that is less sensitive to noise and less ambiguous than sensors such as EMG or HR, which is why it is easier to analyse the results obtained [61].

Along the same lines, eyetracking provides information on eye movement and fixation. Although information can be extrapolated, fixations are not exactly equivalent to attention, and thus it is not always possible to be sure of the results [60]. The reliability of the results can be problematic in some cases.

However, the continued development and improvement of physiological monitoring devices in terms of signal complexity and analysis could mean the increased use of these devices in the field of UX studies in general. With advances in technology, physiological devices have the potential to become more accessible for use in UX laboratories. With more frequent use due to the availability of the technology, more UX professionals could experiment and continue to find new and useful applications of the technology in different sectors. Along these lines, the increasing use of smartwatches that include physiological sensors [62] show considerable promise as a source of data to measure and understand UX during interactions in both real-world and laboratory settings. As Maia & Furtado [9] pointed out, the level of invasiveness of physiological tools is critical, and the level of invasiveness offered by smartwatches is lower than most devices.

Hence, an opportunity opens up in the evaluation of UX through the emergence and combination of different physiological tools that can provide more accurate and reliable results. This would provide a better understanding of the context in which the interaction takes place. In addition, it would be interesting to complement UX evaluation studies with traditional tools such as questionnaires or interviews, which provide a subjective insight into user perceptions. The combination of such different methods would allow for a more complete assessment of UX, i.e. an assessment that includes psychophysiological measures and those obtained through subjective perceptions.

5.2.- THE SAMPLE SIZE, THE KEY ISSUE

Works such as Virzi [63], Lewis [64], and Nielsen [65] claim that 5 participants are sufficient to identify 80% of usability problems. However, the use of small sample sizes can lead to high variability in test results that cannot be fully adjusted [66]. In the study by Faulkner [67] 60 users were analysed and random sets of 5 or more were sampled from the set to demonstrate the risks of using only 5 participants and the benefits of using more. Some of the randomly selected sets of 5 participants found 99% of the problems; other sets found only 55%. With 10 users, the lowest percentage of problems revealed by a set increased to 80%, and with 20 users, to 95%.

In this study, we have tried to find out how professionals perform UX tests with physiological monitoring to obtain an estimate of the theoretically appropriate size. However, the results show that there is no widely validated sample size. Just as usability tests have a limited sample size that is widely discussed by the scientific community, tests that include physiological devices and measure physiological responses do not have a validated sample. A large variability has been observed in the sample sizes of the identified case studies. In this sense, a new research opportunity is identified: to establish an approximate sample size according to the physiological tool to be used in the different sectors.

6.- CONCLUSIONS

User testing is a suitable way to evaluate the UX of a system according to the requirements defined by Väänänen-Vainio-Mattila, Roto, & Hassenzahl [68]. In addition, including physiological tools can help to make a more objective evaluation and provide information without retrospective bias, as the data is collected without interrupting the interaction between the user and the interface. This is an emerging field of study. However, the use of physiological monitoring in UX testing also has its limitations in terms of price, complexity, and time needed to ensure that the evaluation is done properly, when compared to traditional methods. Physiological signals also require a certain degree of interpretation, as the output must be processed to move from raw data to actionable insights.

This review has identified 33 cases of UX studies carried out with physiological monitoring to date. The interest in incorporating psychophysiological measures in UX tests is growing, as shown by the results obtained: 67% of the articles identified in this study have been published in the last 5 years. Furthermore, the high impact factor of the journals shows the acceptance and interest of the scientific community.

EEG has proved to be the most widely used tool. The wide range of studies identified in which this tool is used indicates that its application is increasing in UX testing and that it has great potential for extracting knowledge of emotions despite its complex application protocols and interpretation.

The results of the review show that although the field of app or web solutions is the UX field in which physiological monitoring has been used the most, more and more sectors are appreciating the benefits and possibilities offered by these types of devices. In fact, in the field of app or web solutions, 62.5% of the case studies date from 2018 to 2020 (the year in which the search was conducted). In the case of Gaming, half of the tests identified date from the last five years. In Industry 4.0, only two case studies have been identified from 2018 and 2020. In this sense, it can be considered a field with a great future projection.

In the industrial context, the use of physiological measurements can be beneficial, on the one hand, to obtain objective data on the operator-machine interaction and to understand the complex relationship. On the other hand, to obtain information that can allow the design of solutions that have a positive impact on the operator and, consequently, on the production process.

Finally, this review has allowed us to identify two research opportunities: 1) The evaluation of UX by combining different methods, tools and physiological devices, since the interaction would be evaluated taking into account different types of information. In this sense, physiological devices would provide quantitative and objective data about the user at the moment of the interaction; and questionnaires would obtain qualitative and subjective data, useful to understand user perception of the product. 2) Just as usability tests have a limited sample size, it would be interesting to validate a specific sample size for UX tests carried out with physiological monitoring, depending on the tool to be used.

7.- ABBREVIATIONS TABLE

Abbreviation	Meaning
UX	User Experience
SLR	Systematic Literature Review
HCI	Human Computer Interaction
GSR	Galvanic Skin Response
EDA	Electrodermal Activity
HR	Heart Rate
HRV	Heart Rate Variability
EEG	Electroencephalogram
ECG	Electrocardiogram
BVP	Blood Volume Pulse
EMG	Electromyography

ST	Skin Temperature
BPM	Breathing rate
PPG	Photoplethysmography

Table 12: Abbreviations table.

REFERENCES

[1] ISO 9241-210, "Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems," *International Standard*, vol. 2.Udgave. 2019.

[2] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Applying the user experience questionnaire (UEQ) in different evaluation scenarios," in *International Conference of Design, User Experience, and Usability*, 2014, pp. 383–392, doi: https://doi.org/10.1007/978-3-319-07668-3_37.

[3] A. P. O. S. Vermeeren, E. L.-C. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila, "User experience evaluation methods: current state and development needs," in *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*, 2010, pp. 521–530, doi: https://doi.org/10.1145/1868914.1868973.

[4] D. Wallach, J. Conrad, and T. Steimle, "The UX metrics table: A missing artifact," in *International Conference of Design, User Experience, and Usability*, 2017, pp. 507–517, doi: https://doi.org/10.1007/978-3-319-58634-2_37.

[5] I. Díaz-Oreiro, G. López, L. Quesada, and L. A. Guerrero, "Standardized questionnaires for user experience evaluation: A systematic literature review," in *Multidisciplinary Digital Publishing Institute Proceedings*, 2019, vol. 31, no. 1, p. undefined, doi: https://doi.org/10.3390/proceedings2019031014.

[6] A. Bai and K. S. Fuglerud, "Method for Semi-Automated Evaluation of User Experience Using Brain Activity," *Studies in health technology and informatics*, vol. 256, pp. 811–820, 2018, doi: https://doi.org/10.3233/978-1-61499-923-2-811.

[7] D. L. Neumann and H. R. Westbury, "The psychophysiological measurement of empathy," *Psychology of empathy*, pp. 119–142, 2011.

[8] I. I. P. Oliveira and N. M. Guimarães, "Practical Neurophysiological Analysis of Readability as a Usability Dimension," in *International Conference on Human Factors in Computing and Informatics*, 2013, pp. 194–211, doi: https://doi.org/10.1007/978-3-642-39062-3_12.

[9] C. L. B. Maia and E. S. Furtado, "A study about psychophysiological measures in user experience monitoring and evaluation," in *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, 2016, pp. 1–9, doi: https://doi.org/10.1145/3033701.3033708.

[10] J. J. Braithwaite, D. Derrick, G. Watson, R. Jones, and M. Rowe, "A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments." Accessed: Jul. 13, 2020. [Online]. Available: https://www.birmingham.ac.uk/Documents/college-les/psych/saal/guide-electrodermal-activity.pdf.

[11] "All About Heart Rate (Pulse)." https://www.heart.org/HEARTORG/Conditions/Mo%20re/MyHeartandStrokeNews/All-About-Heart-Rate-%20Pulse_UCM_438850_Article.jsp (accessed Mar. 30, 2021).

[12] J. L. Andreassi, *Psychophysiology: Human behavior & physiological response*. Psychology Press, 2013.

[13] E. Peper, R. Harvey, I.-M. Lin, H. Tylova, and D. Moss, "Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony?," *Biofeedback*, vol. 35, no. 2, 2007.

[14] J. Hopkins, "Electromyography (EMG) | Johns Hopkins Medicine." https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electromyography-emg (accessed Jul. 13, 2020).

[15] C. L. B. Maia and E. S. Furtado, "A systematic review about user experience evaluation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9746, pp. 445–455, doi: https://doi.org/10.1007/978-3-319-40409-7_42.

[16] Taffese, "A Review of Using EEG and EMG Psychophysiological Measurements in User Experience Research," 2017.

[17] J. Robinson, C. Lanius, and R. Weber, "The past, present, and future of UX empirical research," *Communication Design Quarterly*, vol. 5, no. 3, pp. 10–23, Feb. 2018, doi: https://doi.org/10.1145/3188173.3188175.

[18] J. A. Bargas-Avila and K. Hornbæk, "Old wine in new bottles or novel challenges? A critical analysis of empirical studies of User Experience," in *Conference on Human Factors in Computing Systems - Proceedings*, 2011, pp. 2689–2698, doi: https://doi.org/10.1145/1978942.1979336.

[19] M. Z. Baig and M. Kavakli, "A Survey on Psycho-Physiological Analysis & Measurement Methods in Multimodal Systems," *Multimodal Technologies and Interaction*, vol. 3, no. 2, p. 37, May 2019, doi: https://doi.org/10.3390/mti3020037.

[20] D. Tranfield, D. Denyer, and P. Smart, "Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review," *British Journal of Management*, vol. 14, no. 3, pp. 207–222, Sep. 2003, doi: https://doi.org/10.1111/1467-8551.00375.

[21] S. K. Boell and D. Cecez-Kecmanovic, "On being 'systematic' in literature reviews," *Formulating Research Methods for Information Systems: Volume 2*, pp. 48–78, Jan. 2016, doi: https://doi.org/10.1057/9781137509888_3.

[22] R. Caiazza *et al.*, "Business model innovation in the agri-food sector: A literature review," *British Food Journal*, 2016, doi: https://doi.org/10.1108/BFJ-08-2015-0293.

[23] J. Kerr and N. Coviello, "Formation and Constitution of Effectual Networks: A Systematic Review and Synthesis," *International Journal of Management Reviews*, vol. 21, no. 3, pp. 370–397, Jul. 2019, doi: https://doi.org/10.1111/ijmr.12194.

[24] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering - A systematic literature review," *Information and Software Technology*, vol. 51, no. 1. Elsevier, pp. 7–15, Jan. 01, 2009, doi: https://doi.org/10.1016/j.infsof.2010.03.006.

[25] S. R. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 4, pp. 241–257, Nov. 2006, doi: https://doi.org/10.1145/1201730.1201731.

[26] C. Zickler, S. Halder, S. C. Kleih, C. Herbert, and A. Kübler, "Brain painting: Usability testing according to the user-centered design in end users with severe motor paralysis," *Artificial Intelligence in Medicine*, vol. 59, no. 2, pp. 99–110, Oct. 2013, doi: https://doi.org/10.1016/j.artmed.2013.08.003.

[27] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, no. C, pp. 139–183, Jan. 1988, doi: https://doi.org/10.1016/S0166-4115(08)62386-9.

[28] L. Demers, R. Weiss-Lambrou, and B. Ska, "Quebec User Evaluation of Satisfaction with assistive Technology versione 2.0," *The Institute for Matching Persons and Technology: Webster, NY*, 2000.

[29] O. Alzoubi, S. K. D'Mello, and R. A. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 298–310, 2012, doi: https://doi.org/10.1109/T-AFFC.2012.4.

[30] E. M. Holz, J. Höhne, P. Staiger-Sälzer, M. Tangermann, and A. Kübler, "Brain-computer interface controlled gaming: Evaluation of usability by severely motor restricted end-users," *Artificial Intelligence in Medicine*, vol. 59, no. 2, pp. 111–120, Oct. 2013, doi: https://doi.org/10.1016/j.artmed.2013.08.001.

[31] D. Giakoumis, D. Tzovaras, K. Moustakas, and G. Hassapis, "Automatic recognition of boredom in video games using novel biosignal moment-based features," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 119–133, Jul. 2011, doi: https://doi.org/10.1109/T-AFFC.2011.4.

[32] K. Gupta, G. A. Lee, and M. Billinghurst, "Do you see what i see? the effect of gaze tracking on task space remote collaboration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 11, pp. 2413–2422, Nov. 2016, doi: https://doi.org/10.1109/TVCG.2016.2593778.

[33] Y. Liu, J. M. Ritchie, T. Lim, Z. Kosmadoudi, A. Sivanathan, and R. C. W. Sung, "A fuzzy psycho-physiological approach to enable the understanding of an engineer's affect status during CAD activities," *CAD Computer Aided Design*, vol. 54, pp. 19–38, Sep. 2014, doi: https://doi.org/10.1016/j.cad.2013.10.007.

[34] M. Slater *et al.*, "Analysis of physiological responses to a social situation in an immersive virtual environment," in *Presence: Teleoperators and Virtual Environments*, Oct. 2006, vol. 15, no. 5, pp. 553–569, doi: https://doi.org/10.1162/pres.15.5.553.

[35] Y. Bian *et al.*, "A framework for physiological indicators of flow in VR games: construction and preliminary evaluation," *Personal and Ubiquitous Computing*, vol. 20, no. 5, pp. 821–832, Oct. 2016, doi: https://doi.org/10.1007/s00779-016-0953-5.

[36] A. Stein, Y. Yotam, R. Puzis, G. Shani, and M. Taieb-Maimon, "EEG-triggered dynamic difficulty adjustment for multiplayer games," *Entertainment Computing*, vol. 25, pp. 14–25, Mar. 2018, doi: https://doi.org/10.1016/j.entcom.2017.11.003.

[37] J. M. López-Gil, C. Navarro-Molina, R. García, and R. Aleixandre-Benavent, "Análisis de la arquitectura de webs mediante tests de estrés de navegación, de usabilidad y eye tracking," *Profesional de la Informacion*, vol. 19, no. 4, pp. 359–367, Jul. 2010, doi: https://doi.org/10.3145/epi.2010.jul.04.

[38] P. Foglia, M. Zanda, and I. Trading, "Towards relating physiological signals to usability metrics: a case study with a web avatar," *WSEAS Transactions on Computers*, vol. 13, p. 624, 2014.

[39] O. Barral, I. Kosunen, and G. Jacucci, "No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment," *ACM Transactions on Computer-Human Interaction*, vol. 24, no. 6, pp. 1–29, Dec. 2017, doi: https://doi.org/10.1145/3157730.

[40] J. Zhou, S. Z. Arshad, X. Wang, Z. Li, D. D. Feng, and F. Chen, "End-User Development for Interactive Data Analytics: Uncertainty, Correlation and User Confidence," *IEEE Transactions on Affective Computing*, Jul. 2017, doi: https://doi.org/10.1109/TAFFC.2017.2723402.

[41] R. Menges, C. Kumar, and S. Staab, "Improving user experience of eye tracking-based interaction: Introspecting and adapting interfaces," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 6, Oct. 2019, doi: https://doi.org/10.1145/3338844.

[42] K. S. Niksirat, C. Silpasuwanchai, P. Cheng, and X. Ren, "Attention regulation framework: Designing self-regulated mindfulness technologies," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 6, Oct. 2019, doi: https://doi.org/10.1145/3359593.

[43] R. S. Mangion, L. Garg, G. Garg, and O. Falzon, "Emotional Testing on Facebook's User Experience," *IEEE Access*, vol. 8, pp. 58250–58259, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2981418.

[44] V. Meza-Kubo, A. L. Morán, I. Carrillo, G. Galindo, and E. García-Canseco, "Assessing the user experience of older adults using a neural network trained to recognize emotions from brain signals," *Journal of Biomedical Informatics*, vol. 62, pp. 202–209, Aug. 2016, doi: https://doi.org/10.1016/j.jbi.2016.07.004.

[45] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, "EngageMon," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–27, Mar. 2018, doi: https://doi.org/10.1145/3191745.

[46] S. Katsigiannis, R. Willis, and N. Ramzan, "A QoE and Simulator Sickness Evaluation of a Smart-Exercise-Bike Virtual Reality System via User Feedback and Physiological Signals," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 1, pp. 119–127, Feb. 2019, doi: https://doi.org/10.1109/TCE.2018.2879065.

[47] Y. C. Lee, W. C. Lin, F. Y. Cherng, and L. W. Ko, "A Visual Attention Monitor Based on Steady-State Visual Evoked Potential," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 3, pp. 399–408, Mar. 2016, doi: https://doi.org/10.1109/TNSRE.2015.2501378.

[48] J. J. Cabibihan and S. S. Chauhan, "Physiological Responses to Affective Tele-Touch during Induced Emotional Stimuli," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 108–118, Jan. 2017, doi: https://doi.org/10.1109/TAFFC.2015.2509985.

[49] K. Gramann, J. Onton, D. Riccobon, H. J. Mueller, S. Bardins, and S. Makeig, "Human brain dynamics accompanying use of egocentric and allocentric reference frames during navigation," *Journal of Cognitive Neuroscience*, vol. 22, no. 12, pp. 2836–2849, Dec. 2010, doi: https://doi.org/10.1162/jocn.2009.21369.

[50] S. Yan *et al.*, "Exploring audience response in performing arts with a brain-adaptive digital performance system," *ACM Transactions on Interactive Intelligent Systems*, vol. 7, no. 4, pp. 1–28, Nov. 2017, doi: https://doi.org/10.1145/3009974.

[51] A. B. Randolph and M. M. M. Jackson, "Assessing fit of nontraditional assistive technologies," *ACM Transactions on Accessible Computing*, vol. 2, no. 4, pp. 1–31, Jun. 2010, doi: https://doi.org/10.1145/1786774.1786777.

[52] M. Peruzzini, F. Grandi, and M. Pellicciari, "Exploring the potential of Operator 4.0 interface and monitoring," *Computers and Industrial Engineering*, vol. 139, p. 105600, Jan. 2020, doi: https://doi.org/10.1016/j.cie.2018.12.047.

[53] M. Peruzzini, F. Grandi, and M. Pellicciari, "How to analyse the workers' experience in integrated product-process design," *Journal of Industrial Information Integration*, vol. 12, pp. 31–46, Dec. 2018, doi: https://doi.org/10.1016/j.jii.2018.06.002.

[54] P. Sanches, K. Höök, C. Sas, and A. Ståhl, "Ambiguity as a resource to inform proto-practices: The case of skin conductance," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 4, pp. 1–32, Jul. 2019, doi: https://doi.org/10.1145/3318143.

[55] S. M. Cincuegrani, S. Jordà, and A. Väljamäe, "Physiopucks: Increasing user motivation by combining tangible and implicit physiological interaction," *ACM Transactions on Computer-Human Interaction*, vol. 23, no. 1, pp. 1–22, Feb. 2016, doi: https://doi.org/10.1145/2838732.

[56] N. Nourbakhsh, F. Chen, Y. Wang, and R. A. Calvo, "Detecting users' cognitive load by galvanic skin response with affective interference," *ACM Transactions on Interactive Intelligent Systems*, vol. 7, no. 3, pp. 1–20, Sep. 2017, doi: https://doi.org/10.1145/2960413.

[57] S. Jiang, Z. Li, P. Zhou, and M. Li, "Memento: An emotion-driven lifelogging system with wearables," *ACM Transactions on Sensor Networks*, vol. 15, no. 1, pp. 1–23, Jan. 2019, doi: https://doi.org/10.1145/3281630.

[58] F. Karegar, J. S. Pettersson, and S. Fischer-Hübner, "The dilemma of user engagement in privacy notices: Effects of interaction modes and habituation on user attention," *ACM Transactions on Privacy and Security*, vol. 23, no. 1, pp. 1–38, Feb. 2020, doi: https://doi.org/10.1145/3372296.

[59] D. Thakkar, "Biometric Devices: Cost, Types and Comparative Analysis," 2018. https://www.bayometric.com/biometric-devices-cost/ (accessed Jun. 18, 2020).

[60] D. Teverovsky, "A Critical Review of Eye Tracking in UX." Accessed: Mar. 30, 2021. [Online].

[61] A. Bruun, "It's not complicated: A study of non-specialists analyzing GSR sensor data to detect UX related events," in *ACM International Conference Proceeding Series*, Sep. 2018, pp. 170–183, doi: https://doi.org/10.1145/3240167.3240183.

[62] R. Rawassizadeh, B. A. Price, and M. Petre, "Wearables: Has The age of smartwatches finally arrived?," *Communications of the ACM*, vol. 58, no. 1. Association for Computing Machinery, pp. 45–47, Jan. 01, 2015, doi: https://doi.org/10.1145/2629633.

[63] R. A. Virzi, "Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 34, no. 4, pp. 457–468, Aug. 1992, doi: https://doi.org/10.1177/001872089203400407.

[64] J. R. Lewis, "Sample sizes for usability studies: Additional considerations," *Human factors*, vol. 36, no. 2, pp. 368–378, 1994, doi: https://doi.org/10.1177/001872089403600215.

[65] J. Nielsen, "Why you only need to test with 5 users." Useit. com Alertbox, 2000.

[66] A. Cazañas, A. de San Miguel, and E. Parra, "Estimación del tamaño de la muestra para pruebas de usabilidad," *Enfoque UTE*, vol. 8, no. 1, pp. 172–185, 2017, doi: http://dx.doi.org/10.29019/enfoqueute.v8n1.126.

[67] L. Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," in *Behavior Research Methods, Instruments, and Computers*, 2003, vol. 35, no. 3, pp. 379–383, doi: https://doi.org/10.3758/BF03195514.

[68] K. Väänänen-Vainio-Mattila, V. Roto, and M. Hassenzahl, "Towards practical user experience evaluation methods," *Meaningful measures: Valid useful user experience measurement (VUUM)*, pp. 19–22, 2008.