Understanding HCI Practices and Challenges of Experiment Reporting with Brain Signals: Towards Reproducibility and Reuse

FELIX PUTZE, SUSANNE PUTZE, and MERLE SAGEHORN, University of Bremen, Germany
CHRISTOPHER MICEK and ERIN T. SOLOVEY, WPI, USA

In human-computer interaction (HCI), there has been a push towards open science, but to date, this has not happened consistently for HCI research utilizing brain signals due to unclear guidelines to support reuse and reproduction. To understand existing practices in the field, this paper examines 110 publications, exploring domains, applications, modalities, mental states and processes, and more. This analysis reveals variance in how authors report experiments, which creates challenges to understand, reproduce, and build on that research. It then describes an overarching experiment model that provides a formal structure for reporting HCI research with brain signals, including definitions, terminology, categories, and examples for each aspect. Multiple distinct reporting styles were identified through factor analysis and tied to different types of research. The paper concludes with recommendations and discusses future challenges. This creates actionable items from the abstract model and empirical observations to make HCI research with brain signals more reproducible and reusable.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **HCI theory, concepts and models;**

Additional Key Words and Phrases: Brain sensing, electroencephalography, EEG, functional near-infrared spectroscopy, fNIRS, experiment model, reproducibility

ACM Reference format:
Felix Putze, Susanne Putze, Merle Sagehorn, Christopher Micek, and Erin T. Solovey. 2022. Understanding HCI Practices and Challenges of Experiment Reporting with Brain Signals: Towards Reproducibility and Reuse. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 31 (March 2022), 43 pages.
https://doi.org/10.1145/3490554

1 INTRODUCTION & MOTIVATION

In **human-computer interaction** (**HCI**), there has been growing interest in integrating brain-sensing into interactive systems. The field is developing traction and we anticipate further expansion as sensor technology becomes more practical for interactive applications. While HCI

The work is partially supported by the U.S. National Science Foundation under Grant No. 1835307 and the DFG project "DINCO - Detection of Interaction Competencies and Obstacles" under Grant No. 316930318.
Authors' addresses: F. Putze, S. Putze, and M. Sagehorn, University of Bremen, Bibliothekstraße 1, 28359, Bremen, Germany; emails: {felix.putze, sputze}@uni-bremen.de, mmsagehorn@web.de; C. Micek and E. T. Solovey, Worcester Polytechnic Institute (WPI), 100 Institute Road, Worcester, Massachusetts, 01609-2280; emails: {cjmicek, esolovey}@wpi.edu.

1073-0516/2022/03-ART31
https://doi.org/10.1145/3490554

researchers have demonstrated many novel ways of integrating brain data into HCI practice, there are several challenges that slow progress toward reaching the potential for real-world use.

In particular, HCI research with brain signals requires a diverse skill set, including understanding of brain function, signal acquisition methods, signal processing, and machine learning, in addition to HCI, design, and software engineering. Each of these aspects is themselves active research area. This steep learning curve and interdependence of numerous research areas limit the pace of innovation. To overcome this challenge, researchers often attempt to build on each other's work. However, this has been difficult to date, and it is more common to see each research group creating their own datasets, analysis pipelines, and following their own methodologies to move their research forward.

A contributing factor to the limited research reuse is the differing norms stemming from research cultures in related fields, making it challenging to publish, reuse, and reproduce work in HCI venues. Reviewers often have contradictory expectations, depending on their background. In addition, page limitations and cultural norms result invaluable details omitted that could support the reproduction, reuse, and extension of published HCI research.

As an example, we consider the creation of adaptive interactive systems that modify system behavior based on a cognitive or affective state, measured by brain sensing. Such systems require the development of real-time capable technology for processing of brain input (including the necessary infrastructure to capture, transmit, and process the data), along with an experiment paradigm for collecting an appropriate experimental dataset to work with. They also require the implementation and fine-tuning of user interface components and a strategy to adapt to brain data, as well as the integration of the components in an application, and finally, some form of user evaluation.

Currently, researchers start from scratch when implementing and studying such systems. This is not only very costly and time-consuming, but also creates opportunities for errors and requires extensive expertise in a number of different fields. Under such conditions, reproducible research and sharing of resources and datasets are especially important for the effective progression of the field. Once the research is under review in HCI venues, some reviewers may expect rigorous experimental controls as in neuroscience, while other reviewers may a expect clear demonstration of usability, and still, others may value highly novel and visionary proof-of-concept demonstrations over rigorous experiments. In HCI, contributions can be made to the field from many of these perspectives, and more, but it can be challenging to understand what is valued.

To move towards better reproducibility and sharing of resources, there are two related gaps that need to be filled. First, there is a need for systematic exploration of the approaches, methods, and applications in HCI research with brain signals today, to identify common themes and potential for re-use. This involves identifying recurring domains and applications, types of modeled cognitive and neurological states, methods of integrating brain signals in HCI research, and the method and sensor for measuring brain activity. Secondly, it is important to study the current practices for presenting and documenting work in this research area to ensure the feasibility of reproduction and reuse. Both of these goals require the collection and curation of a large database of HCI publications that specifically focus on the use of brain signals in HCI research.

1.1 Contributions

This article addresses the gaps described above, providing a foundation for conducting reproducible and reusable HCI research with brain signal data. Specifically, this article reports on:

(1) The curation of a dataset of 110 HCI articles published since 1996 that explore brain sensing. Through analysis of these published works, we can look at the current state of HCI research and reporting practices when working with brain signals.

(2) Systematic analysis of the dataset to compare domains, applications, modalities, measured mental states, and processes. *This work substantiates existing manifestations of HCI research with brain-sensing today and identifies commonalities as well as heterogeneity within this emerging, independent research area.*
(3) Coming from the broad variety of research identified in (2), we created an overarching experiment model that provides a formal structure for describing HCI research with brain signals. The experiment model serves as a starting point for further analysis on how experiments in this field are reported and what aspects may be of relevance (See (4)). The model provides definitions as well as examples for each aspect and is evaluated by a group of external experts. *This step creates an abstract foundation of experiment reporting beyond individual articles.*
(4) We compare the experiment model with the reality of the curated dataset of HCI articles. We perform statistical analysis to investigate the structure of experiments and experiment reporting in a data-driven way. *This provides an evaluation of the model as well as analysis of reporting structure and reporting gaps, and reveals connections to the different sub-groups identified in (2).*
(5) Informed by our data-driven analysis of current practices within and outside of HCI, we derive a set of recommendations and considerations for future authors and reviewers and discuss future challenges. *This creates actionable items from the abstract model and the empirical observations in the analysis.*

2 RELATED WORK

Multiple processes have been developed that support the goal of reproducible research, such as citable open data repositories and pre-registration (such as the Open Science Framework[1] or Zenodo[2]). More and more, publication outlets and funding agencies expect the publication of recorded data to ensure that other researchers can check the reported results, but also benefit from the investment of time and money. It should be noted that open data (in the sense of releasing purely the raw recorded data) on its own is not enough to ensure replicable research: If the data is not well-documented, the ability of independent researchers to reproduce or build on it is restricted. This challenge in the curation process of open data is related to the *long tail of science*. This term describes the phenomenon that most data is produced in small, individual research projects with specific applications, domains, or restrictions in mind (see [42] for an analysis of this phenomenon in the field of neuroscience). Making this data accessible to other researchers requires a joint understanding, or even better, *a formal model* of how the data should be shared and what information is necessary to use the data successfully.

Disciplines like genomics [64], systems biology [146], enzyme activities [135], and materials science [113] have developed explicit models which allow the description of experiments in a common language on the basis of common attributes, enabling re-use of data. An experiment model provides a list of mandatory or optional attributes, each of which describes a specific aspect of an experiment. Attributes are structured in categories of related items. Often, these models are associated with software ecosystems that support the documentation, publication, and querying of data. For example, STRENDA DB offers an experiment model to describe research data on enzyme activities with a minimum amount of information using 15 attributes describing the enzyme itself, the determination method, and the results [135]. Even such a minimal experiment model was shown to prevent common documentation gaps in the literature [53]. Consequently, over 50 publishers, e.g., *Nature* and *PLoS*, already recommend using such a model as part of the

[1] https://osf.io/.
[2] https://zenodo.org/.

publication workflow.[3] More domain-independently, O'Connor et al. [100] suggested a generic model to describe metadata associated with scientific experiments with semantic descriptors.

In other fields, some of which are closely related, there are standard practices in place to facilitate reproduction and reuse of prior work, building a solid foundation that later research can build on. For example, the field of neuroscience has invested much effort into the creation of standards for published datasets with a joint documentation model, such as the OpenNeuro initiative.[4] Assessment of the impact of shared brain imaging data on the scientific literature has found that sharing data can "increase the scale of scientific studies conducted by data contributors, and can recruit scientists from a broader range of disciplines" [89]. Babaei et al. [9] extracted published practices of EDA recording and analysis and published a structured and modifiable version of the resulting model[5] and similarly, Bergström et al. [13] analyzed reporting strategies for Virtual Reality experiments and distilled a checklist for future publications. For structured publication and documentation of experimental processes, approaches like open electronic lab notebooks (e.g., "Open Lab Book"[6]) have emerged, offering this opportunity. In the field of machine learning, which is often applied to brain signals, it has become increasingly common to publish executable code on the basis of open toolkits and programming languages, often on version control platforms such as GitHub. This overcomes the difficulty of reporting all algorithm parameters in a page-limited article. The "Articles with Code" database[7] provides a repository for publications with associated code.

There also exist unifying initiatives in the **brain-computer interfaces (BCI)** community, which has some overlapping as well as some clearly distinct goals from the HCI community. The BNCI Horizon 2020 initiative [17] is a large-scale effort for standardization in conceptualizing and communicating of BCI technology. A difference to the approach in our article is that the BNCI initiative is mostly driven by the BCI community, and focuses strongly on aspects of signal processing and machine learning. We will see later that many applications of brain signals in HCI are not BCI-related (i.e., do not involve real-time processing of brain activity) and we also argue that the HCI community of researchers using brain signals overlaps with the BCI community to some extent, but is ultimately different from it and exhibits different needs for publication. The BNCI initiative outlines four main challenges, including the lack of a "common terminology" and the absence of "curated benchmark datasets." It also acknowledges that "the BCI community might not be fully aware of relevant work in other fields," highlighting the need for communication in interdisciplinary formats, which are a staple of the HCI community. In the context of this initiative, there also exists a database[8] of shared BCI datasets, all documented in a relatively homogeneous format, with a focus on the data structure and semantics. In one of the few conceptual works in the field, Kosmyna and Lécuyer [68] establish a conceptual framework to characterize different works using brain signals along four axes (some with sub-axes) related, for example, to the type of input or the underlying neural mechanism. Many other reviews in BCI literature summarize parts of the research landscape systematically: For example, Aricò et al. [8] discussed passive BCIs that are developed for use outside the lab, and Lotte et al. [82] explored BCI in combination with **Virtual Reality (VR)** technology. While these reviews give a good overview on the state-of-the-art methods, they do not have a particular HCI focus and do not discuss experiment conduction and reporting. In contrast, a recent monograph provides an introduction and review of brain input research from an HCI perspective [129], but is not centered on experiment reporting,

[3] https://www.beilstein-institut.de/en/projects/strenda/journals/.
[4] https://openneuro.org/.
[5] https://edaguidelines.github.io/.
[6] https://openlabnotebooks.org/.
[7] https://paperswithcode.com/.
[8] http://bnci-horizon-2020.eu/database/datasets.

reproducibility, and reuse. In comparison to past publications, our article focuses on a tailored subgroup of articles, namely HCI-centric work which uses brain data. This focus gives us the opportunity to dive deeper into the specific characteristics of these articles and their reporting practices. However, this does not mean that the other meta-analyses in BCI are not applicable; indeed, we find that many of their observations regarding good research practices are a common thread which we also encounter here.

The **Association for Computing Machinery** (**ACM**) has introduced an alternative approach to strict enforcement of particular practices. It instead uses positive reward mechanisms through the concept of badges that reward the publication of artifacts, such as code or data, for reproducing research results.[9] This approach has seen adaption at some ACM venues, such as ACM RecSys,[10] however, it is not yet widespread at HCI conferences. We think that is because HCI is not a purely computational discipline, but a field that combines computational, experimental, analytical, and human-centered aspects; thus, it may not be clear what information is needed and expected for reproducible research.

Still, efforts toward reproducible research and an open data culture have reached the HCI community as well and have sparked discussions about the utility and necessity of replication studies in HCI [57],[11] the benefits and challenges of open source and open data publishing [37], the use of pre-registration [29]. This resulted in changes to incorporate reproducibility and transparency in the review process of HCI conferences, such as ACM CHI.[12] Besides CHI and ACM, several other publishers which are relevant to the HCI community[13] encourage their authors to publish their experiment protocols and refer to them within the corresponding manuscript. Despite such initiatives on an organizational level, Wacharamanotham et al. [143] investigate the status quo in the transparency of CHI research artifacts and come to the conclusion that most researchers do not publish their research artifacts due to misunderstandings about possible options and the reasons behind making them available. Echtler and Häußler come to a similar conclusion with regards to open-source publication of HCI software [37]. We find that in a diverse and interdisciplinary field such as HCI, it is not always clear what this should look like and how it applies to each sub-field within HCI. By examining current practice in the HCI subfield using brain signals, our article takes initial steps toward building this understanding.

3 LITERATURE DATABASE CURATION

In this section, we describe our methodological approach to collecting and curating a large database of publications on the use of brain signals in HCI research. We created the database with two central goals in mind. First, we wanted to identify how the field is structured in terms of domains, employed sensor technology, common brain input paradigms, recurring cognitive and neural constructs, and other attributes. Additionally, we looked at factors that are often tied to reproducibility, such as the number of participants and the number of times the data or artifacts of a article were actually building on another study. The methods and results of this investigation will be presented in Section 4. Second, we wanted to systematically categorize and statistically analyze the current reporting practices, looking for commonalities as well as gaps, with an eye toward reproducibility and reuse of brain signal research in HCI. The methods and results of this investigation will be presented in Section 6.

[9] https://www.acm.org/publications/policies/artifact-review-badging.
[10] https://recsys.acm.org/recsys20/call/#content-tab-1-1-tab.
[11] http://www.replichi.com/.
[12] See https://chi2020.acm.org/blog/changes-to-the-technical-program/#more-1870.
[13] E.g., https://plos.org/open-science/, https://www.hindawi.com/journals/ahci/guidelines/.

Table 1. Counts of Regarded Articles Per Year

Year	Number of Articles	Year	Number of Articles
1996	1	2011	9
2001	1	2012	8
2002	1	2013	7
2003	1	2014	13
2004	3	2015	16
2006	1	2016	11
2007	1	2017	10
2008	2	2018	8
2009	6	2019	8
2010	3		

We considered publications that made use of brain signals in combination with HCI methods. This often falls into two categories: (1) the application of HCI methods to systems employing brain input (e.g., usability evaluation of a brain-based adaptive system), or (2) the use of brain signals to create or enhance HCI methods (e.g., using brain signals to evaluate workload induced by a user interface). In the selection, we aimed at excluding publications that focus solely on the processing of brain signal data to study signal processing or machine learning techniques, without an intention to employ the results to HCI-related research questions. To identify such publications, we included only peer-reviewed articles that were published at dedicated HCI conferences or journals. We systematically examined contributions from the main proceedings (full articles, accessed through the official publication platforms, such as the ACM Digital Library) of the following conferences: ACM CHI, ACM UIST, ACM IUI, ACM ICMI, NordiCHI, ACM CSCW, and ACM UbiComp. As a starting year, we chose the year of the first international CHI conference, which is 1992. The first eligible publication was published in 1996 [138]. Table 1 summarizes the counts of regarded articles per year. Additionally, we considered publications in the following journals: ACM ***Transactions on Computer-Human Interaction*** (**TOCHI**), ***International Journal of Human-Computer Studies*** (**IJHCS**), ***Interaction with Computers*** (**IWC**), and ***Frontiers on Human-Media Interaction*** (**FOHMI**).

For a article from any of these venues to be included, it must contain at least one of the search terms: **electroencephalography** (**EEG**), **functional near-infrared spectroscopy** (**fNIRS**), MRI, brain, or BCI (or variants, such as Brain-Computer Interface). In total, we considered publications from 125 proceedings and journal volumes. From this initial selection, we excluded the following entries: (1) Any conference publications that were not considered full articles (e.g., extended abstracts), (2) articles for which the search term occurred in a different context than anticipated (e.g., in a article about the "brainstorming" technique), and (3) articles for which the search terms occurred only in the introduction, related work, or future work (e.g., articles which discuss brain input as an alternative or extension to their presented work). After applying these exclusion criteria, we ended up with 110 articles eligible for the survey. The created database is available as supplementary material to the article.

We are aware that the selection of venues excludes some potentially relevant articles from the analysis, for example, ones focusing on BCI itself, neuroergonomics, or specific domains to which brain signals can be relevant. While such works can be very important from an HCI perspective, identifying them among other, less HCI-oriented articles in said fields, is very difficult without formal filtering criteria. We made this decision to ensure that the authors themselves considered HCI to be the central contribution in their work, thus minimizing the chance of adding "false

positives" to the database. By publishing research at an HCI venue, authors signal that they consider their work to be most impactful in the field of HCI. They also use this decision to select a specific target audience, namely the HCI community for which brain input is one tool among many, in comparison to a venue dedicated to such systems. As these conferences and journals employ HCI experts as reviewers and editors, we can assume that the accepted manuscripts indeed have a focus on HCI. Additionally, the selected venues all put an explicit focus on interdisciplinary research, which makes it more likely that the publications addresses a wider audience. The fact that other HCI-oriented research on the use of brain data exists does not limit the generalizability of our results as we still cover a sample of 110 articles with an in-depth analysis.

4 CHARACTERIZING HCI ARTICLES WITH BRAIN SIGNALS

To gain a better understanding of what HCI research contributions with brain signals look like today, we systematically explored the curated literature database. We were interested in characterizing the diversity and common themes across published HCI articles, the relationship to other fields, and the potential for reproducibility and re-use.

4.1 Methods

For each article, we collected a number of "article demographics" to characterize the state of HCI research with brain signals. On the one hand, we were interested in aspects, which provide a general context of the article to study different types of investigations. On the other hand, we looked at aspects related to the reproducibility of the reported research. These demographics include:

- *Application/Domain*: For what domain or specific application is the brain signal used? (e.g., education, robotics)
- *Cognitive/affective state or process*: Measured brain activity is usually inspected for specific patterns related to specific cognitive or affective states or processes. (e.g., error potentials, cognitive workload)
- *Brain signal integration type*: These are described in more detail at the end of this section, and include explicit control, implicit closed loop, implicit open loop, neurofeedback, mental state assessment, and HCI evaluation.
- *Brain signal modality*: Which type of signal is used to capture brain activity? (e.g., EEG, fNIRS)
- *Main contribution*: From the abstract, we extracted the sentence that summarizes the main contribution of the article.
- *Number of participants*: We included this item because a frequent point of discussion in reviews and studies on reproducibility is the number of participants in an experiment and the resulting power of the analysis.
- *Is follow-up*: Is this article a conceptual or concrete follow-up to an earlier publication of the same or other authors?
- *Exploratory or Confirmatory*: We look for evidence for whether the analysis is *confirmatory*, i.e., driven by explicit *a priori* hypotheses, or *exploratory* analysis to generate new hypotheses.

The first four categories were inspired by an analysis of the 218 unique keywords authors assigned to their articles. Apart from a small number of specific methodological keywords referring to the type of analysis employed in the article, all keywords could be grouped into these categories. It is interesting to note that only four generic keywords "BCI" (42 times), "EEG" (49), "fNIRS" (17), and "human-computer interaction" (13) occur more than 10 times, showing the large variety of topics covered by these articles.

The employed definitions for the types of *brain signal integration* are based on the characterization presented by Krol et al. [74], extended by additional categories which play an important role in HCI (*neurofeedback* and *evaluation*).[14] The categories can first be differentiated in *online* and *offline* use cases of brain input. The online ones are then further differentiated by how the measured brain activity is exploited. In particular, some systems use *explicit* or *direct* control and others use *implicit* or *passive* input [32, 136, 153, 154]. With *implicit* input, brain signals that occur naturally are detected passively in real-time, with no special effort from the user. Unlike most *explicit* control systems, which often utilize brain data as the primary system input, *implicit* input paradigms frequently integrate brain signals as a secondary input channel to interactive systems. Implicit paradigms can further be broken down into the *open loop* and *closed loop* systems. Thus, the categories that we used for integration type are described below.

- **Explicit Control:** Also referred to as *direct control*, paradigms in which mental commands (such as motor imagery, directing the attention towards a blinking pattern) are given by the user and these are mapped directly to user interface operations (e.g., cursor movement or letter selection). These are sometimes categorized into *active control* and *reactive control* paradigms [155], depending on whether brain activity is controlled independent of any external stimuli or in response to external stimuli, respectively.
- **Implicit Closed Loop:** The initiated response to specific aspects of the measured brain activity is designed to directly influence these aspects through *online* system behavior changes. For example, a closed loop system could measure workload and adapt the interface specifically to adjust the user workload.
- **Implicit Open Loop:** *Online* use of brain input in which specific cognitive or affective states or processes from the data is detected and used to inform and adapt an interactive system. Open loop refers to the absence of any direct or intended coupling of the adaptation back to the input. For example, when detecting changes in workload, an open loop system may notify the user.
- **Neurofeedback:** *Online* use of brain signals which visualizes or otherwise backchannels aspects of the captured neural activity (in raw or processed form) to the user, enabling them to self-regulate their own brain activity consciously (in contrast to processing the brain activity to adjust the interface as for implicit closed loop).
- **Mental State Assessment:** *Offline* processing of brain sensor data for the classification of specific cognitive or affective states or processes from the data. This analysis is performed as a self-contained methodological contribution, without leveraging the result further. The goal is often to transfer the results to an online system at a later point.
- **HCI Evaluation:** *Offline* processing of brain sensor data with the explicit purpose of evaluating stimuli or a (non-BCI) HCI system offline from brain signal data. While the previous types of brain signal integration describe HCI approaches that employ a brain input component, evaluation is concerned with the use of brain signals to analyze general HCI systems. In contrast to the mental state assessment, this is not done with the intention of explicit integration in the system control loop and usually focuses on low-level neural responses in contrast to high-level cognitive states.

4.2 Results

The earliest article that we identified is from 1996 [138], outlining a vision of using brain signals among other methods as a control signal for human-computer interfaces. It took five years for the

[14] [74] also defines the concept of "Automated Adaptation". We did not encounter any article that qualifies for this kind of system during our analysis.

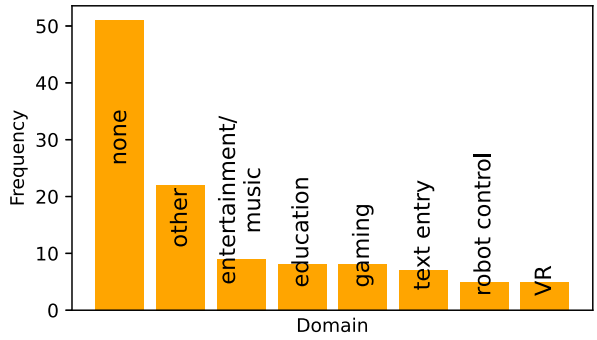

Fig. 1. Frequency of domains/applications in analyzed articles.

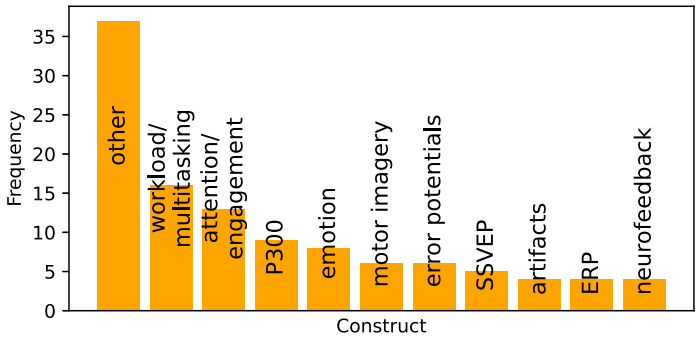

Fig. 2. Frequency of mental states or processes modeled (only entries with more than three articles are shown).

next article [35] to appear and five more for the third one [77]. Since then, the number of articles related to brain signals in HCI increased steadily over the years. Of all 110 articles, most of them (42) were published at the ACM CHI conference, which is also the largest venue, and which has seen continuous articles on the topic each year since 2008. Below, we discuss some findings.

4.2.1 Applications and Domains. We were interested in exploring applications and domains in which HCI research with brain data was situated. Figure 1 illustrates that there is no single dominant domain or application in which brain signals are used for HCI. Even the most frequent domain, "entertainment/music," occurs in only nine publications, followed by "education" and "gaming." These domains are followed by three recurring application types, namely "text entry," " VR," and "robot control." All other domains occurred three or fewer times. However, 41 publications are not rooted in a domain or specific type of application at all. This shows on the one hand the universality of brain signals in HCI, but also hints at a current lack of grounding in real-world, ecologically valid scenarios.

4.2.2 Cognitive/Affective State or Process. To better characterize the different ways of utilizing brain signals, we look at the constructs which are measured and modeled in the articles. The most frequent ones are presented in Figure 2. Here, we make the observation that different types of reporting exist. Some authors frame the modeled constructs from a neurology-centered perspective, referring to characteristic neural responses such as P300, or **steady-state visually evoked potentials** (**SSVEP**). Other authors frame the modeled brain activity from a cognition-centered perspective, referring to processes and states such as workload, attention, and so on. It must be

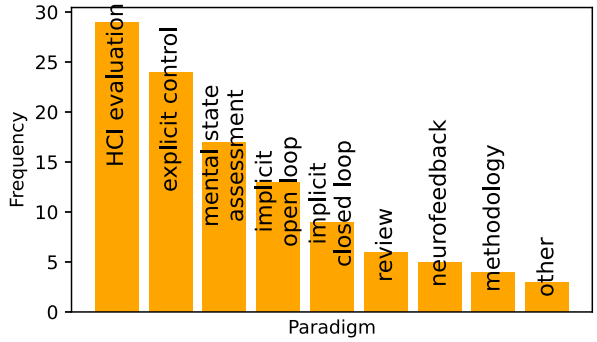

Fig. 3. Frequency of types of brain signal integration.

stressed that both perspectives can refer to the same phenomenon: For example, a P300 is a neural response to a stimulus that can be explicitly modeled and studied, but it can also be considered as a correlate of an attention process. We feel that ultimately, it is important to be specific from both perspectives, as only that guarantees a validity of the measurement, as well as a useful interpretation of it in an HCI context. Also, by providing both perspectives, there are broader possibilities for re-use by other researchers. The analysis still reveals the existence of recurring states and processes, most notably cognitive workload and attention. Besides, several other "one-off" processes exist, such as "humor" [12], "presence" (in virtual reality) [133], or "expertise" [31] (all combined in the "other" category of Figure 2). This indicates the brain signal is employed for a large variety of purposes and thus has a large potential of impacting different areas of HCI. A consequence is that it is more difficult as a community to establish common datasets, baselines, and feature sets.

In contrast to narrow, precisely defined low-level neural effects like the P300, high-level cognitive concepts are broader and encompass several aspects. This can mean that cognitive or affective states, which are actually different (such as stress and workload [120]), are clustered together as the classifier does not allow a more granular representation. Thus, two articles both concerned with a construct like "attention" should not be treated immediately as measuring the same aspect. Fairclough [41] points out that the relationship between measurements and target states (or between low-level neural features and high-level cognitive processes) is not necessarily a one-to-one relationship. Thus, an interpretation for HCI purposes needs to take such ambiguity into account and needs to validate newly defined constructs.

4.2.3 Brain Signal Integration Types. Another important aspect of the existing HCI research is the type of brain signal integration. This provides insight into the ways brain data is integrated into HCI research. Figure 3 gives an overview of the frequency of the different categories. Additionally, Table 2 gives an overview of all regarded articles and how we categorized them according to integration type. Open loop and closed loop systems, as systems, which adapt to the user's state, come closest to the vision explicated by Velichkovsky and Hansen [138] in their 1996 CHI article. Indeed, they appear regularly in our analysis. However, we also see that even if considered together, these categories do not form the most frequent type of HCI publication with brain signals. One reason is that an open or closed loop system requires expertise in recording and processing neural signals, real-time processing and classification of the data, integration into a non-trivial application, as well as experimental paradigms to train and test the systems under realistic conditions. Consequently, a considerable number of publications concentrate only on parts of this chain, such as the mental state assessment, but leave out others, such as the actual testing of a quantifiable usability improvement (these constitute the majority of articles in the "mental state assessment"

Table 2. Categorizations of All Articles in the Dataset According to the Input Integration Type

Input Integration Type	Articles
HCI evaluation	[56], [55], [93], [66], [102], [79], [3], [5], [63], [91], [19], [118], [83], [26], [44], [31], [92], [132], [49], [14], [71], [12], [25], [78], [47], [142], [133], [134], [28],
explicit control	[35], [72], [119], [145], [151], [101], [99], [114], [51], [52], [33], [150], [98], [106], [76], [90], [70], [69], [103], [39], [84], [65], [95], [36],
implicit open loop	[121], [131], [108], [58], [140], [122], [2], [107], [7], [109], [27], [73], [112],
implicit closed loop	[128], [126], [130], [1], [111], [152], [147], [115], [15],
neurofeedback	[45], [43], [54], [81], [6],
mental state assessment	[138], [77], [50], [23], [154], [61], [157], [139], [110], [125], [11], [88], [34], [148], [48], [158], [96],
other	[127], [123], [104], [85], [21], [136], [87], [41], [38], [10], [4], [24], [97],

category). Systems which are able to bring all components together show that using brain signals in runtime can yield substantial usability improvements [1, 152] or unlock completely novel kinds of applications [103].

The largest individual category is "HCI Evaluation." Here, brain signals are used as a tool to evaluate or compare specific aspects of the application, such as the workload induced by the user interface or the degree of distraction by certain sonifications. The benefit of using brain signals for HCI evaluation is that they are continuous, non-interrupting, and objective measures, which can be hard to measure otherwise. HCI evaluation approaches often apply brain signals offline, i.e., they do not require real-time capabilities and single-trial classification. This lowers the bar for bringing brain signals to productive use compared to real-time systems. Of all articles which use brain signals in real-time, the majority follow an "explicit control" paradigm, i.e., translating explicit mental commands into control commands or text. Here, we observe that these articles more often than not regard a brain-computer interface as the subject of study, which is evaluated with BCI methods.

4.2.4 Brain Signal Modality. From a technical standpoint, the oldest technology for BCI, EEG, is still the most prevalent (see Figure 4), as 78% of all articles use EEG as a brain input modality. As a regularly recurring contender is fNIRS, appearing in 16% of all articles. Three articles employ MRI for brain imaging. While hybrid EEG/fNIRS interfaces have been a staple approach in more BCI-oriented research, none of these works have yet found their way into HCI publications.

When taking a closer look at the employed EEG headsets, we can differentiate between "research-grade" and consumer-grade headsets. The latter group is mainly comprised by the NeuroSky Mindwave, the Epoc Emotiv, and the InteraXon Muse. Characteristics of consumer-grade headsets are a low cost, short setup time, no requirement for electrode gel, and high comfort, but also a limited number of electrodes in a fixed montage. There has been an ongoing debate on the signal quality and susceptibility to artifacts when using consumer-grade devices, where some authors see a drastic reduction in their validity [18, 116], while others claim them to yield useful neural data [117, 156], or arrive at mixed conclusions [86]. The prevalence of consumer-grade headsets is significantly higher (following a McNemar test, $p = 0.001$) for articles, which are rooted in some kind of application compared to those without. This shows that the benefits of consumer-grade headsets outweigh their drawbacks in such situations. Figure 5 breaks down headset types by the brain signal integration paradigm. This analysis shows that consumer-grade headsets are most

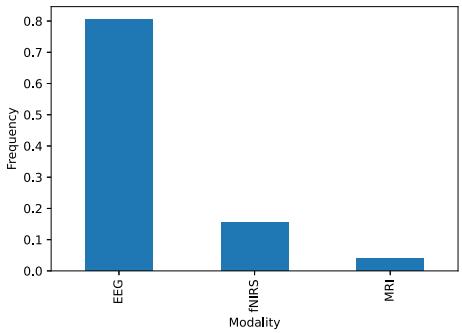

Fig. 4. Relative frequency of brain measurement modality.

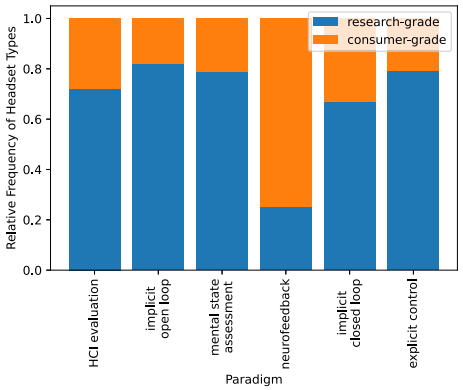

Fig. 5. Relative frequency of research-grade and consumer-grade EEG headsets for different brain signal integration paradigms.

prevalent in neurofeedback applications, where minor measurement errors might go unnoticed by the user.

A consequence of the omnipresence of artifacts in neural data is that we always need to consider that detected effects could result from such artifacts and not actual neural processes; i.e., validation of effects is an important part of the analysis. In complex HCI applications, especially outside the laboratory, a clear discrimination of neural activity and artifacts is often difficult. This also means that many published research results which employ brain data in uncontrolled HCI scenarios without a validation of the underlying neural signals (e.g., when using the output of commercial EEG appliance as a blackbox) need to be taken with a grain of salt. However, there is a different perspective to this argument. In their summary of a workshop on BCI in the wild, Friedman, et al. [46] conclude that while "From a basic research perspective it is essential to distinguish between information extracted from the brain and other types of information picked up by the brain sensors", however "From a practical point of view, however, [they] believe there is no reason to limit ourselves to "pure" brain interactions". This seems to be one of the major breaking points between the fields of BCI and HCI using neural signals (see also Section 4.2.7). Note that this relaxation of expected "purity" in HCI should not void the aspiration to extract at least some actual neural information from the recorded data; otherwise, the researchers would be better off recording muscle activity, eye movement, and so on. directly with more appropriate sensors.

4.2.5 Contribution Types. By analyzing the self-reported main contribution from the abstracts of the articles, we categorized articles as technical, user-centered, or theoretical contributions. As a technical contribution, we counted any contribution which was measured using a metric directly calculated from the brain signal, such as classification performance or significant differences in neural activation between different conditions. 56.5% of all contributions fall into this category. As a user-centered contributions, we consider contributions that are measured with metrics related to the user experience, such as satisfaction or task efficiency. 39% of all articles fall in this category. As a theoretical contribution, we count any work which is non-empirical, such as a review, or a conceptual work. Only 2% of all contributions can be considered theoretical.

4.2.6 Number of Participants. One important issue of empirical research and an important point in the discussion on reproducibility is about the number of participants in a study. Traditionally, one would expect a power analysis to estimate the ideal number of participants [149]; however,

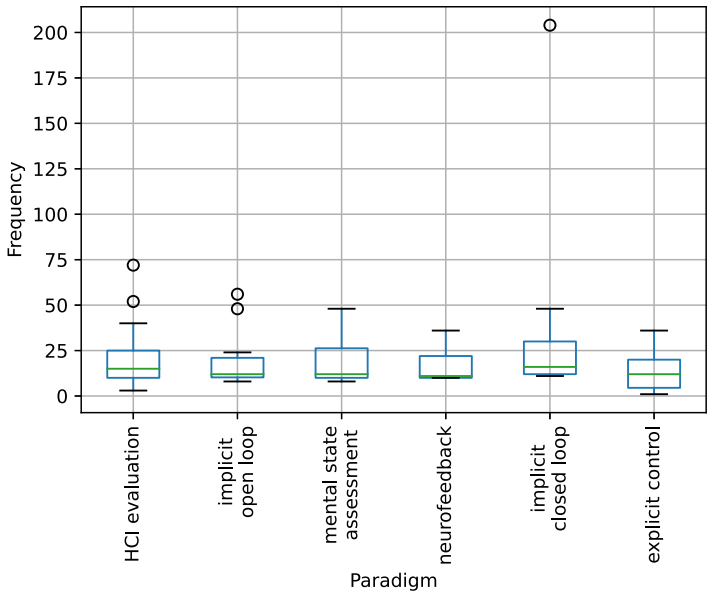

Fig. 6. Number of participants by brain signal integration paradigm.

estimating effect size reliably is difficult when each published system is very different from the others. In the analyzed articles, an average of 20.1 participants appears in the studies; however, this number is inflated by a few large studies with more than 100 participants. The median is much lower with 12 participants. This is enough to find large effects (effect size >0.8) under favorable conditions, but also hints at the fact that reproduction and re-use of published results may help to strengthen the results; performing complex experiments with neural input technology is time-consuming and gains little recognition beyond the necessary minimum and distributing this effort across multiple studies from different laboratories could accelerate progress in the field. Figure 6 breaks down the average number of participants for different input paradigms. We see that beyond a few outliers, the number of participants is fairly consistent around the reported average for all paradigms.

4.2.7 Emergence of a Distinct Sub-field. Over the course of this history, the HCI research community using brain signals has developed into its own interdisciplinary sub-field, drawing methods and applications from multiple different technical and human-centered areas. It is noteworthy that this community is connected to, but not identical with the conventional BCI community, which focuses more strongly on clinical applications and technical aspects of using brain signals in real-time systems. To quantify the independence of the fields, we checked (for the years 2017-2019) how many of the authors of our analyzed articles were also authors of articles at some of the most central BCI conferences, the BCI Meeting of the BCI Society, the IEEE Brain–Machine Interaction Workshop, and the Graz BCI Conference. This analysis revealed that fewer than 10% of all authors in our survey also appeared at these venues and fewer than 40% of articles have at least one author appearing. This shows that HCI conferences and journals are not just another publication outlet for BCI researchers. It also implies that the HCI-focused community may have its own goals and culture of performing and evaluating experiments. In addition, publishing work at BCI conferences often has different expectations as the submitted article or abstract is often not as heavily reviewed and not considered an archived publication. It should be noted that this does not influence the

analysis in this section, which only concentrates on the author lists of these publications, not on their content (while the BCI conferences are not the primary publication outlets of that community, the attending researchers can still be considered representative).

To illustrate that these fields (BCI and HCI using brain signals) are distinct from each other, we identified two pairs of articles, each containing one article from each field with a similar topic, and compare how the articles are structured and what kind of information they present (Tables 3 and 4). This comparison of representatives of the respective fields allows us to see typical differences between the two fields in approaching common research questions. As exemplary articles, we picked Szafir and Mutlu [130] and Myrden and Chau [94] for our first pair and Yuksel et al. [151] and Krusienski et al. [75] for the second. We chose these articles as (i) each pair has the same expressed goals (of modeling a person's level of attention from EEG to improve subsequent interactions and improving P300-based BCI systems, respectively), (ii) they are all reasonably well-cited (254 vs. 44; 35 vs. 774) and published in typical outlets for the respective fields (ACM CHI vs. IEEE Transactions on Neural Systems and Rehabilitation Engineering; ACM CHI vs. the Journal of Neuroscience Methods), and (iii) are comparable in length (word count: 8797 vs. 11711; 2818 vs. 4239). For brevity, we denote the articles by HCI1 and BCI1; and HCI2 and BCI2, respectively. We discuss differences between the articles in six different areas: (a) Motivation, (b) Setting, (c) Construct differentiation, (d) Signal Processing, (e) Metrics of success, and (f) Validation approaches. Tables 3 and 4 contrast these areas for the two article pairs.

In summary, we find that BCI1 and BCI2 go deeper into the neural origin of the data and how signal processing impacts classification performance. In contrast, HCI1 and HCI2 go deeper into the user experience outcomes and a provide greater context of the application and the impact thereon.

4.2.8 Follow-ups, Replication, and Reuse. Of all analyzed articles, only 19% were characterized by the authors explicitly as follow-up, articles to already published works within or outside our analyzed dataset. As follow-up, we counted any work which directly linked to previous work conceptually, replicated an experiment setup or re-used artifacts such as an application or processing code. In most (but not all) cases, this was a direct continuation of previous research of the same group. We did not find any publications that directly distributed their data or their software artifacts. This does not mean that these are not available at all (they might have been published at a later point in time or are available on request). Both observations indicate the difficulty of building on existing work.

4.2.9 Exploratory & Confirmatory Analysis. To better understand what kind of research approaches were used in the studied articles, we investigated how many articles were performing *confirmatory* analysis in comparison to *exploratory* studies. This distinction was introduced to mainstream HCI research [29] and related disciplines along with the critique of widespread **"Hypothesizing After the Results are Known"** (**HARKing**), i.e., unwarranted flexibility in statistical analysis to adjust the methods to the observations.) and the proposal of pre-registering research protocols, which requires a clear distinction between confirmatory and exploratory analyses [22, 137]. The difference is particularly relevant when considering follow-ups of existing work; while exploratory research warrants confirmatory replications to consolidate or falsify emerging knowledge, confirmatory research can be considered a foundation for new approaches and experiments. The research and analysis discussed in this article, for example, is exploratory in nature; we study a new research topic on newly collected data, without explicitly pre-defined hypotheses. This implies that follow-up research (by the authors or other parties) should confirm the findings on data from similar fields or future years.

Of all articles in our dataset, none performed a formal pre-registration (and many of these predate the widespread discussion of this concept). Only a few examples explicitly position themselves

Table 3. Comparison of Experiment Representation in Two Articles from HCI and BCI (First Pair)

Hci1 [130]	Bci1 [94]
Motivation	
Targeted "appropriate, effective responses" in "educational settings"	Expressed the goal to "facilitate greater accuracy" of the (active BCI) systems they want to improve, i.e., they aim for a more technical in a more focused area of application.
Setting	
Implemented their model in a humanoid robot that performs real-time adaptation; participant interaction is structured coarsely. Both studies were conducted in a laboratory setting.	Required participants to perform a selection task using a traditional P300 speller.
Construct Differentiation	
Took a relatively broad strokes approach for the definition of attention, explicitly stating that the terms engagement and attention are used "interchangeably".	Differentiates three related, but distinct mental states (fatigue, frustration, and attention) and manipulates them individually.
Signal Processing	
Described a custom mapping algorithm that builds on the "proprietary" preprocessing of the consumer-grade EEG hardware.	Elaborated the parameter values for how their signals were filtered and digitized, but goes on to state that "all aspects of data collection and experimental procedure were controlled" by the BCI system that was used. Choices of electrode placement, reference, and signal decimation were experiment parameters that were manipulated over the course of data collection.
Metrics of Success	
Looked at objective and subjective outcome measures, such as task success rate and attitude towards the agent or motivation.	Assessed and compared classification performance using the accuracy score.
Validation Approaches	
Used a manipulation check to test whether the attention model reacts to different situations. They further look at the participants' objective and subjective responses to the robot behavior and differentiate between genders.	Discussed the neural mechanisms potentially underlying the observed classification performance; the authors also discuss how electrode selection and electrode reduction influence the classification performance.

as being explicitly exploratory (e.g., [134]) or confirmatory. To still operationalize this distinction, we scanned the articles for concepts related to confirmatory analysis, as compiled from and the preregistration guidelines collected by Open Science Framework[15]: explicitly formulated *hypotheses*, pre-defined *independent and dependent variables*, accounting for *multiple testing*, estimating *effect sizes*, and a *power analysis* to determine an optimal sample size. While many other aspects play a role in confirmatory research [144], these were the most important ones which also could be identified objectively. We accounted for common variants in terminology (i.e., "multiple testing"

[15]https://osf.io/zab38/wiki/home/.

Table 4. Comparison of Experiment Representation in Two Articles from HCI and BCI (Second Pair)

HCI2 [151]	BCI2 [75]
Motivation	
Developed a "multi-touch table using [a] P300-based BCI" to explore "the embedding of BCI in new HCI situations"	Seeked to "improve classification performance" of a traditional P300 speller
Setting	
Participants performed analogous selection tasks using both a traditional P300 speller and a new "multi-touch system" that relied on a multi-touch table to detect the presence of objects and flash the areas beneath them	Required participants to perform a selection task using a traditional P300 speller.
Both studies used within-subjects, trial-based experiment designs, and performed both offline and online classification.	
Construct Differentiation	
Does not greatly differ across the two articles; both describe how participants must "attend" to a target stimulus in order to select it with their respective P300 systems, but do not attempt to specifically define or measure "attention."	
Signal Processing	
Used MATLAB software provided by the BCI headset vendor to handle signal acquisition and classification.	Elaborated the parameter values for how their signals were filtered and digitized, but goes on to state that "all aspects of data collection and experimental procedure were controlled" by the BCI system that was used. Choices of electrode placement, reference, and signal decimation were experiment parameters that were manipulated over the course of data collection.
Metrics of Success	
Sought to show that participants who used their multi-touch P300 system selection success that was comparable to a traditional P300 speller	Aimed at determining the set of recording and classification parameters that resulted in optimal performance for a standard P300 speller paradigm.
Validation Approaches	
Compared the online classification performance of their multi-touch P300 system to a P300 speller, as well as to speller performance from a prior recent study	Analyzed offline classifier accuracy in a factorial fashion across multiple recording and feature configurations to determine the optimal set of parameters, then validated their choice with another, online classification experiment.

and "[Bonferroni|...] correction"), excluded any uses of these terms which were clearly false positives (i.e., when using the term "correction" in reference to self-correcting interfaces), and removed nine articles from the analysis that do not contain any original empirical research. As a result, we found that 27.7% of all articles defined hypotheses explicitly, and 18.8% do so for (in)dependent variables. 21.8% of articles perform a correction for multiple testing. 5% of articles mention effect sizes (all post-hoc when discussing the results), and only one is a statistical power analysis. This

result shows the openness for innovation and exploration within the HCI community, but also hints at the fact that the field needs to consolidate its knowledge in a more systematic way.

The same method allows us to study the usage of statistical methods more generally: We find that 93% of articles use basic descriptive statistical methods (keywords: *mean, average, median, mode, sd, standard deviation, standard error, maxim(um|a), minim(um|a), correlation, accurac(y|ies)*). Beyond that, 76% of articles use some kind of statistical modeling (keywords: *linear model, nonlinear model, Gaussian mixture model, factor analysis, pca, test, score, p (<|=), anova*).

5 EXPERIMENT MODEL FOR HCI RESEARCH WITH BRAIN SIGNALS

In examining articles covering a broad variety of use cases, it becomes apparent that different publications contain different approaches of reporting the experiments, even within the field of HCI. One aspect may be covered in great detail in one article and only briefly touched upon in another. Sometimes, this is due to the type of research contribution (e.g., whether it involves the application of machine learning techniques), while other times it is due to cultural differences (e.g., method-driven vs. application-driven research).

Thus, to investigate how HCI experiments involving brain signals are described in the literature, we first define a *model* of such experiments. The benefit of a model is that it provides vocabulary and structure for comparing different works and reporting styles and can act as a guideline on what other researchers report and expect to be reported. An experiment model is a superset of items, or *attributes*, occurring multiple times in the relevant publications, which is grouped into different categories. Each attribute defines a unit of information related to the experiment. Not all attributes are necessarily applicable to every article. However, they should be general enough to appear regularly and across a number of applications and use cases.

The model we discuss here was created in an iterative process during initial passes of the surveyed literature: A superset of reported attributes was created, similar concepts (or identical concepts with different names) were grouped together and categories were defined in accordance to the typical section structure of the articles. This process resulted in the following *categories*: Technical aspects of recording, Task description, Participants, Experiment flow, Data processing, and Brain signal integration. Of course, such an experiment model contains aspects, which may also be covered in a general model of experiments in HCI or cognitive psychology; however, we opted for creating a model specific to brain signals in HCI to cover the unique combination of experimental work in an HCI context together with aspects of cognition, neuroscience, signal processing, and machine learning. In the future, the model could be extended to cover other types of physiological interfaces.

In Table 5, we introduce these categories and list the attributes contained within them. Tables 6–11, in the Appendix, provide further details, including definitions and examples. For each attribute, we give a definition of the item and present an example taken from the surveyed literature. We tried to identify examples, which are biased towards a more detailed documentation, but individual examples may still lack certain information.

6 REPORTING PRACTICES IN HCI RESEARCH WITH BRAIN SIGNALS

In this section, we discuss our analysis to determine what aspects of an experiment the HCI community considered relevant to report, what structure we could identify through statistical analysis (compared to the a-priori structure of the experiment model), and what differences we encounter in relation to the variance of research as observed in Section 4. For this purpose, we annotated each of the 110 articles, as described in the Section 3 section according to the presence or absence of each attribute in the presented experiment model (Section 5). Thus, we not only extracted statistics

Table 5. Summary of the Experiment Model, Listing All Regarded Attributes of HCI Experiments Involving Brain Signals. Attributes are structured in several categories

Category	Experiment Attributes
Technical aspects of recording	type of sensor, sensor position, sampling rate, measurement quality, reference, auxiliary signals, synchronization with stimuli and other signals, recording environment
Task description	participant restraints, output devices, input devices, middleware/communication, framework/technical platform, task functionality, architecture, stimulus material, visualization provided?, timing, code for task provided?
Participants	recruitment strategy, incentives, age, gender, occupation, inclusion or exclusion criteria, approval of ethics committee
Experiment flow	experiment structure, instructions, training procedure, trial ordering, repetitions, blocks & breaks, pre-study screenings, questionnaires
Data processing	derivation of labels, data transformation, filtering, windowing, artifact cleaning, hyperparameter optimization, outlier handling, feature extraction, feature selection, learning model, evaluation procedure, processing code provided?
Brain signal integration	brain input effect, type of integration

Tables 6–11, in the Appendix, provide further details, including definitions and examples.

on *what* was investigated (Section 4), but also *how* it is reported. We then explore the prevalence of different attributes across articles and the relationships between different attributes and topics.

The goal of this review was not to judge specific articles for the absence or presence of certain details, but to get a picture of how experiments are reported today, and what differences we observe in the style of reporting for different types of articles. This provides an indication of what aspects of the experiment model are central to most HCI publications, what parts need refinement, and what parts are considered niche.

Attributes could be rated as a present, absent, or not eligible (for example, "brain input effect" for non-real-time studies). Every article was examined by one to four raters (median is 2), primarily authors of the article and research assistants, all with a background in human-computer interaction, computer science and/or neuroscience, and including graduate students, a senior postdoctoral researcher, and a professor. We first compiled a set of detailed definitions and examples for the different attributes and then all raters annotated the same article and discussed the process to identify any questions or areas that needed additional clarity. Between raters, we arrive at an agreement rate of 81%. All raters could add additional information (e.g., additional attributes) or refinements to the definitions of the model during the annotation process. Ambiguities could be settled through a growing collection of examples and explanations. The annotated list of all articles as well as the data matrix of ratings is available as supplementary material to this article.

6.1 Prevalence of Experiment Model Attributes

In Figure 7, we show the prevalence of the attributes of the experiment model across articles, i.e., the relative number of articles reporting the corresponding aspect. We see a large variety of resulting scores: A number of attributes related to data acquisition are present in nearly every article in which they are eligible: These include mostly basic signal-related aspects, such as the type of device/sensor or sensor positioning. On the other end of the spectrum, we find that many attributes of data processing and machine learning are rarely reported, even if eligible.

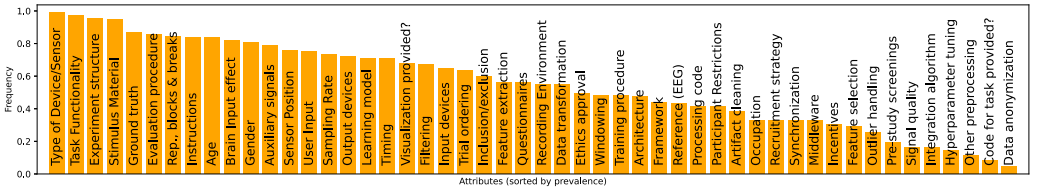

Fig. 7. Prevalence of attributes across all analyzed articles, values indicate the relative number of articles reporting the corresponding aspect.

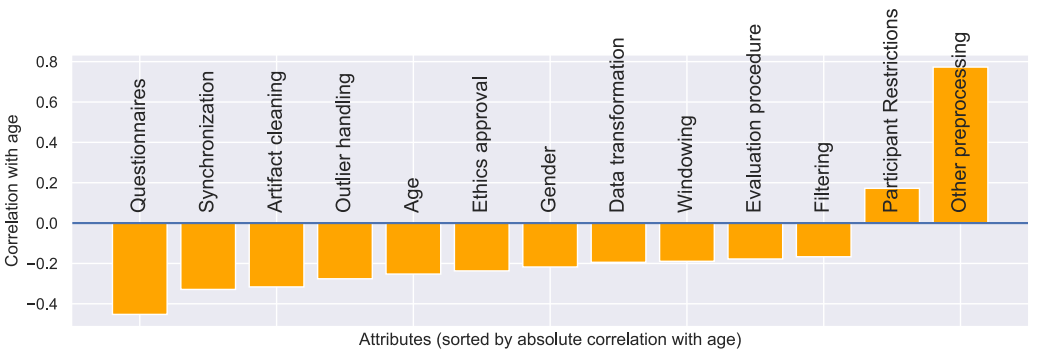

Fig. 8. Correlation of model attributes prevalence and article age for selected items, sorted by correlation coefficient.

One point to consider is that we rated attributes as absent in cases where no information was given, even if logically possible, because it was not considered by the authors. An example of such a case is the attribute "participant restraints". If the authors did not apply any restraints and thus do not report on this, this is indistinguishable from a situation in which a restraint was applied but not reported. A consequence of this observation is that it may be necessary to explicitly report attributes of the model as not applicable, to avoid such ambiguity.

To get a rough estimate of the impact of reporting these attributes, we calculated the Pearson correlation between the fraction of reported attributes, which were present in a article with the number of citations per year since publication. This yields a modest, but significant correlation of $r = 0.16$ ($p = 0.04$), indicating that comprehensive experiment reporting may play some role in the article's impact and success. Similarly, we correlated the number of reported attributes with the age of an article, to study whether reporting changes over time globally. The correlation of $r = -0.12$ is not significant ($p = 0.12$). For a more detailed view, we repeated this analysis for individual attributes and found that some attributes, such as questionnaires and methods of synchronization, appear more frequently in recent publications (occurrence of these attributes correlates negatively with age), while others, such as participant restrictions or non-standard preprocessing techniques, occur less frequently in newer publications (occurrence of these attributes correlates positively with age). See Figure 8 for the attributes with the highest absolute values for r.

6.2 Structure of Experiment Reporting

In the following, we analyze the collected prevalence information on experiment model attributes to uncover a structure. While the proposed experiment model in Section 5 is structured based on expert perspective, we also want to investigate a data-driven approach to uncover a structure in the reported experiment aspects from their occurrence in different articles.

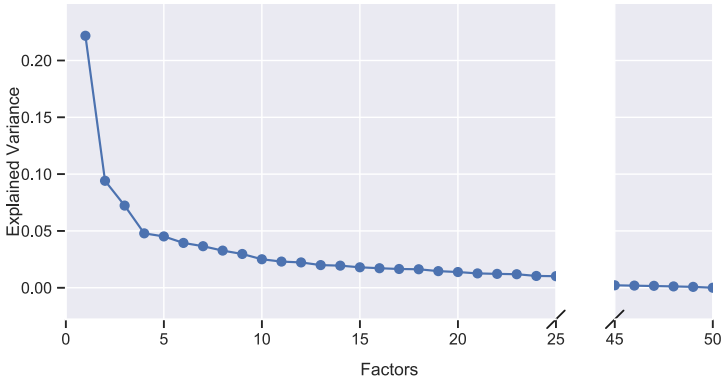

Fig. 9. Scree plot of variance explained by each factor, resulting from the factor analysis on experiment model attributes. Based on the leveling slope after factor 10, we decide for this number of factors for further analysis.

As a first step towards this goal, we investigate the (in)dependence of the different attributes of the experiment model, to detect aspects that often occur combined and may thus identify a common category of attributes, which are associated with a specific type of research. Additionally, this analysis allows us to identify any potential redundancy in the attributes. We calculated the Pearson correlation between all pairs of attributes across all scores. If r^i_{max} is the maximum absolute correlation coefficient of attributes i with all other attributes, then the average r^i_{max} across all i is 0.56, indicating that on average, the different attributes can indeed be considered sufficiently distinct. The pair with the highest correlation of 0.89 is "gender" and "age" from the category "participants". Still, this correlation analysis showed that patterns of jointly occurring model aspects exist.

We exploited this fact to uncover the structure underlying the analyzed articles, potentially revealing different reporting patterns for different types of publications. For this purpose, we performed an explorative factor analysis with Varimax rotation on all attributes. The factor analysis reveals hidden *factors*. Each factor groups multiple related attributes of the model together. As the number of factors is usually smaller than the number of attributes, the analysis reduces the dimensionality of the data, while preserving as much variance as possible. The factor analysis therefore, allowed us to identify, through a data-driven process, a structure in the form of groups of experiment-reporting attributes that occur jointly in many articles. To preserve the variance of the original data, we avoided any scaling of variables or other preprocessing which could influence it.

Figure 9 shows a scree plot of the variance associated with each factor, representing the explained variance for each resulting factor (factors are ordered by variance in descending order). This analysis helps us to understand how many factors we want to retain: Factors with high levels of explained variance are important, interpretable contributors to the overall model, while later factors explain less and less variance and are usually more difficult to interpret. From this result, we decided to continue with ten factors, as the curve flattens out after this point.

Figure 10 shows the loadings of all attributes for the first ten factors. The loading of an attribute for a factor can be understood as the correlation between the attribute and that factor. Therefore, the higher the absolute loading value is, the stronger that attribute is tied to that factor and predicted by it. On the other hand, a loading of close to 0 indicates the absence of a (linear) relationship between the attribute and the factor. This allowed us to find interpretations of these factors. Some of the factors represent well the categories defined by the experiment model. For example, F1 strongly corresponds to the "participants" category. Factor F2 loads highest on

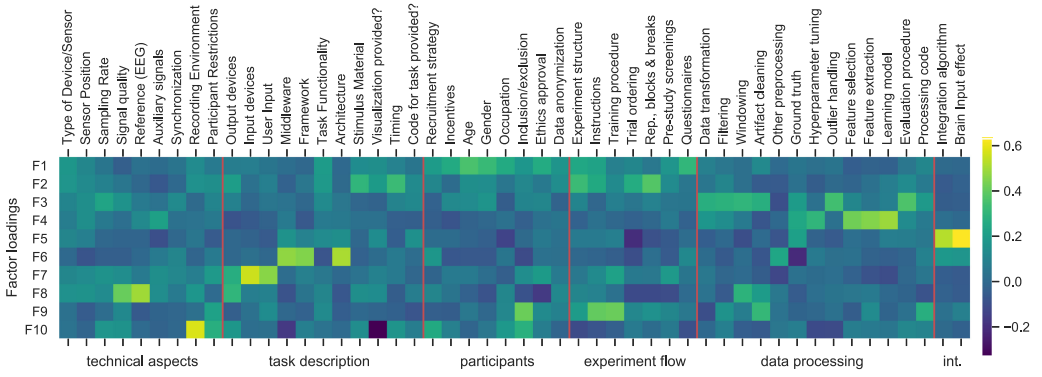

Fig. 10. Loadings of all attributes (larger values, represented in yellow, indicate a high loading) for the first ten factors (F1–F10) in the explorative factor analysis. The order of the attributes reflect the proposed experiment model categories, with red lines indicating the borders between categories of the experiment model. This enables exploration of how each factor relates to the experiment model attributes and categories.

attributes from the "experiment structure" category, however it also contains aspects of other categories, such as "Stimulus Material" (5^{th} highest loading) or "Timing" (7^{th}). Therefore, we interpret this factor as relating to aspects of "Experiment Design". Factors F3 and F4 split up the "data processing" category of the model, by differentiating between general signal processing techniques and attributes exclusive to machine learning approaches. Factor F5 reflects the category of "brain signal integration." Finally, F6 is related to software-related attributes from the technical aspects. It should not be surprising that some factors only cover parts of a category as we consider more factors than categories exist in the experiment model proposal. Further factors then become more specialized.

While we think that the design of the experiment model and the grouping of attributes in categories reflects a logical structure, the result of the factor analysis suggests a possible alternative based on typical use cases. We can formalize these alternative categories in a data-driven way by standardizing all loading vectors to unit length and then assigning each model attribute to the factor on which it is loading highest (given the normalized loadings). This way, the attributes are not grouped by the pre-defined categories but by the data-driven ordering. The result of this process is shown in Figure 11. Based on the attribute loadings of each factor, we assign semantic categories that summarize the associated model aspects: F1: Fundamentals, F2: Experiment Design, F3: Data Processing, F4: Machine Learning, F5: Integration, F6: Software, F7: User Input, F8: Data Acquisition, F9: Quality Assurance, F10: Recording Conditions.

Finally, we investigated whether there are patterns in reporting, based on various aspects of the research. Figure 12 shows the average weight of each factor (F1, …, F10) for articles grouped according to different criteria: type of brain signal integration, type of contribution, and type of headset. The figure shows us that different types of integration set very different foci when reporting their experiments. In some cases, this directly results from the nature of underlying study designs. For example, neurofeedback systems were investigated in controlled experiments less often than other types of integration, thus, loading lower on F2 (Experiment Design, loading −0.94 vs. a mean of −0.11) and by their nature often do not provide other input techniques besides the BCI itself, thus, loading lower on F7 (Input, loading −1.12 vs. a mean of −0.10). Similarly, implicit closed loop systems naturally load highly on F5 (integration, loading 1.33 vs. a mean of 0.16), while mental state assessment articles focus strongly on F4 (machine learning, loading 0.79 vs. a mean of −0.03). In other cases, the difference may result more from a different prioritization. For

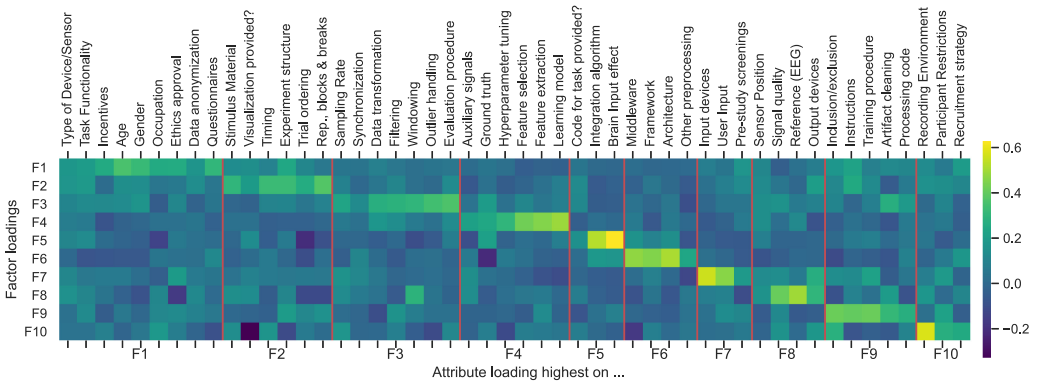

Fig. 11. Reorganization of experiment model attributes, sorted to reflect the assignment to factors, red lines indicate the borders between assigned factors. Loadings of all attributes (larger values, represented in yellow, indicate a high loading) for ten factors (F1–F10).

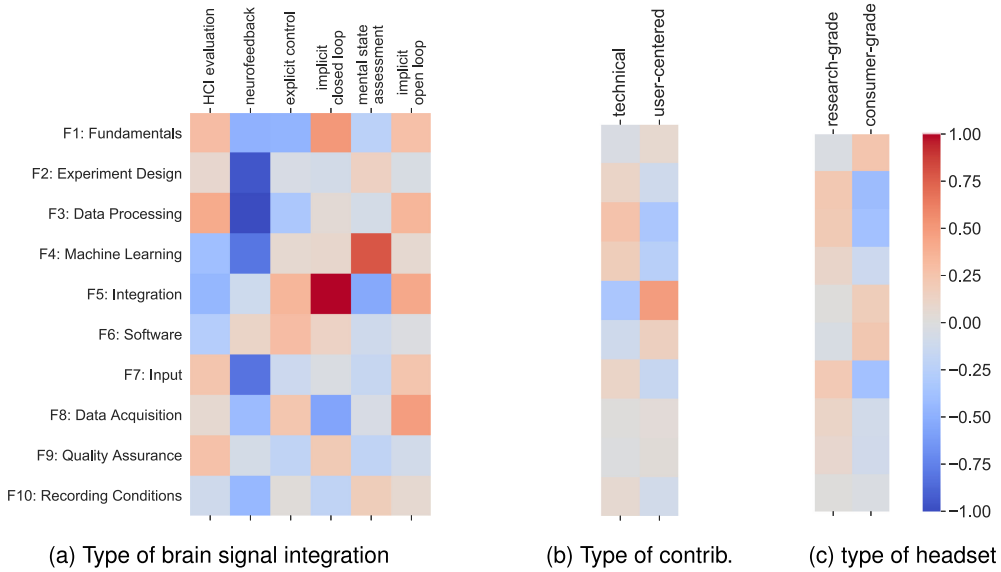

Fig. 12. The average weight of factors for articles grouped according to different criteria, illustrating how strongly (or weakly) different factors of the experiment model are associated with specific groups of articles. A high (low) value indicates that attributes associated with that factor are reported more (less) frequently for that type of article than average.

example, articles of the "HCI evaluation" type provide on average the most detailed description of data processing (F3, loading 0.4 vs. a mean of −0.11), although most of the other articles likely employ similar techniques as well. Similarly, "implicit closed loop" publications seem to put more weight on the documentation of the experiment fundamentals (F1) than other types of integration (loading 0.5 vs. a mean of −0.01). The differences for a type of contribution and type of headset are much less pronounced in comparison. This indicates that different reporting approaches are mostly a result of the type of integration and that these categories thus form a useful high-level categorization.

7 EXPERT PERSPECTIVES AND REVIEW

In Sections 5 and 6, we explored a model for the description of empirical work in using brain input with HCI methods and explored the current practices in the existing literature. To understand what information researchers find to be relevant for reproducing and reusing work, we distributed an online questionnaire to experts and collected their feedback and comments. This is described below.

Questionnaire Structure. First, the questionnaire asked for background information about the participant. Then, it asked questions to understand the extent to which the participant felt that their own articles were reproducible, followed by questions about their own sharing practices. The questionnaire then asked about the participant's experience reproducing or reusing other published work. Finally, the questionnaire contained all of the potential experiment model attributes along with a description and example as in the Appendix. Participants were asked to consider three contexts: *experiment reproduction*, (i.e., re-performing the same experiment), *data reuse* (e.g., running different analysis on published data), and *artifact re-use* (e.g., extending a presented software framework with additional functionality). For each of these three contexts, the participants rated the level of detail needed for each item as one of the following: *Not Needed at All, Not Very Detailed, Moderately Detailed,* or *Highly Detailed*, which have been coded as 0, 1, 2, or 3 respectively. Throughout, there were opportunities for the participants to provide free comments.

Participants. To recruit experts to the study, we contacted all authors of articles considered in the previous sections. We received feedback from 12 external experts (from 6 different countries, ranging from graduate students to full professors from HCI and BCI) in the field who commented on the model and provided suggestions for minor adjustments. Two participants did not complete the entire survey, and we include their ratings and responses for the questions that they did answer.

Data and Code Sharing Practices. Four out of ten of the participants regularly share recorded data, data processing code, experiment processing code, and study materials. Seven out of ten felt that most of their published studies are reproducible from the information given in the article. Space limitations were identified by four participants as the barrier to publishing fully reproducible experiments. In addition, one participant mentioned that the data is often regarded as medical data that has privacy requirements and cannot be shared and another participant mentioned that permission from the participants may not have been given for sharing. Another participant discussed the additional time requirements for preparing data for sharing. One participant mentioned that application or data processing code might not be shared for intellectual property protection.

Experience with Reproducibility and Reuse. We also asked the participants to reflect on any experiences that they have had reproducing study results from published work by other researchers. All six of the participants who indicated that they have attempted this, expressed that doing so was challenging and time-consuming in some way. One mentioned getting lower accuracy than reported in the original article. Another mentioned that lack of documentation about channel positions, markers, and so on, made it difficult. One response directly contrasted the neuroscience/BCI/machine learning journals and the HCI publications, stating that the former is "much more reproducible" than the latter. Specifically, this participant mentioned that the former provides more details on algorithms, parameters, and precise timing of the protocol, while the latter is often, but not always, "vague/generic". They also pointed out the difference between practices of sharing code with "BCI/machine learning articles do[ing] it much more often (but not enough)."

Research Community Views on Reproducibility. We also asked participants to rate the level to which reproducibility is valued in their main research community. Six participants rated it either

4 or 5 out of 5 (high). Of these, one researcher identified their main fields as perceptual graphics, and the other five identified BCI as their main research area. The four other participants rated this as 1 or 2 (not valued highly) and they identified their primary field as human-computer interaction, cognitive systems, or computer security user behavior, with one mentioning it being "valued in theory, but no way of receiving credit for it".

Feedback on Experiment Model Attributes. The participants provided 41 comments to the different model attributes, leading to an incremental improvement of the category definitions. Their responses showed they rated the attributes of the model with an average of 2.04, with no model category scoring lower than 1.92. This indicates that the model lists relevant information. Bonferroni-corrected pairwise Wilcoxon signed-rank tests showed that their importance ratings differed significantly depending on whether the experts rated the attributes for the purpose of *experiment reproduction*, (median importance of 2.26), *data re-use*, (2.12), or *artifact re-use*, (2.0). This shows that the context and viewpoint of the author and reader influence the priority for different attributes.

Summary. Many of the comments about experiences with reproduction and reuse of their own work as well as that of others align and confirms the motivation of this article. In addition, their feedback on the experiment model attributes constitutes a preliminary evaluation of the model.

8 DISCUSSION & OUTLOOK

From our observations in the literature survey and the experiment model, as well as the insights from the expert questionnaire, we derive a number of recommendations.

Exchanging Best Practices

Our analysis showed that different types of articles use different practices and could benefit from the exchange. For example, we showed a divide between application-oriented and method-oriented HCI research in the types of sensor headsets used and the important trade-off between user comfort and signal quality. This divide becomes especially important if we consider that a substantial number of articles do not report details on aspects related to signal processing, sometimes taking the output of the commercial systems at face value. On the other hand, the method-oriented community could learn from others that comfort and user experience matter, and working towards that is a merit which may even warrant a certain loss of classification performance (but not empirical validity). There is an opportunity for these articles, which have different end goals, to be written to have broader appeal and potential for reuse. Another cultural difference lies in the way that the modeled mental states and processes are referred to, i.e., from a neural perspective or a cognitive perspective. A common vocabulary and a better awareness of similarities and relations could help to better leverage common resources and insights.

It would also be valuable to turn specifically to the publications of the core BCI community to transfer some of the relevant standard practices of these experts in experiment design, signal processing, machine learning, and the underlying neural processes to the HCI community, to avoid common pitfalls, especially if they are BCI-specific. In particular, Brouwer, et al. describes six recommendations related to BCI studies, which include (1) define your state of interest and ground truth (2) connect your state of interest to neurophysiology, (3) eliminate or address confounding factors, (4) practice good classification methods, (5) provide insight into classification results and (6) justify the use of neurophysiology [16]. Jeunet, et al. details best practices related to signal acquisition, data processing, experiment design, and the user component [62]. These BCI-focused articles make recommendations that can carry over into HCI research with brain signals.

HCI does not have a consistent record of meta-research and model-building as has helped the BCI field, (e.g., through the BNCI 2020 initiative), to establish common terms and public datasets.

Besides improving internal communication, it would also help to make the field more accessible to HCI researchers who do not yet use brain signals in their research but could benefit from it.

Awareness of the Heterogeneity of HCI Research with Brain Signals

As authors, reviewers, and editors in the field, it is important to be considerate of the heterogeneity of HCI research with brain signals. To study and organize reporting practices in breadth, we first analyzed a range of domains, of cognitive vs. neural states, of types of headsets, types of brain input integration, and so on, to get an understanding of the diversity of concepts a model needed to cover. This "demographic heterogeneity" directly impacts experiment reporting. For example, the choice of headset determines whether details about sensor positions, filtering, data quality can be reported, and the factor analysis reveals differences in reporting for different types of integration.

Our analysis provides some evidence that the HCI research community using brain signals is different from the traditional BCI community. Furthermore, we showed that even within the community, there are different sub-communities with different reporting styles. This observation is not only relevant for future authors seeking guidance, but also for reviewers, who may use our findings to see what level of detail can be expected on the one hand, but also to recognize the range of acceptable contributions, on the other hand. Because of the diverse disciplinary backgrounds in the field, a critical challenge in publishing is the misalignment of expectations (e.g., articles being criticized for being "too neuroscientific" for an HCI conference, or articles "lacking rigor" when using commercial devices). Another source of heterogeneity is the wide-spread in the acceptable number of study participants. Interdisciplinary research then runs into the risk of being "shot down" by multiple disciplinary reviewers, pulling the article in multiple, opposite directions. This may discourage attempts at interdisciplinary work, which we consider crucial for the long-term development of the field. By uncovering the model that represents current HCI research with brain data, based on existing peer-reviewed publications, further discussion can emerge about expectations, reproducibility, and opportunities to strengthen future research in the field.

This recommendation should not be understood as a call for a lack of carefulness or willingness to improve. On the contrary, the diversity in research creates ample opportunity for learning for everybody, as elaborated in the previous recommendation. For example, researchers need to be aware and properly address the frequent problems that can lead to overfitting and non-reproducible results, such as improper handling of artifacts and premature parameter tuning. Utilizing strict validation criteria with pre-defined metrics, and appropriate chronological validation during offline analysis as well as the use of established processing pipelines can prevent that. However, not all learning is actionable immediately, e.g., because data has already been collected or necessary collaborations with experts from other areas need to be build up. We, therefore, think that in many cases, the inclusion of a transparent and honest discussion of shortcomings and opportunity for improvement may be a chance to offset imperfect, but thought-provoking research that others can build on (see also the third recommendation below).

Creating Opportunities to Build on Each Other's Work

Our analysis has shown that no publication or series of publications by the same group of researchers covers all aspects in the maximal depth. A reason for this is that complex research on brain input often involves multiple real-time components for signal processing, machine learning, and user interface design and no individual researcher is an expert in all of these areas. We recommend leveraging the fact that the research community in this area is large (396 unique authors contributed to the articles we analyzed) and diverse. This goal has several implications: First, it is necessary to provide enough information for others to be able to build on the work. Again, the experiment model can help to provide a sufficient level of detail. Second, we also recommend moving

away as a community from expecting articles with full end-to-end systems as well as large-scale user studies all as novel contributions. A publication can have merit if it provides the research community with an exciting new experimental paradigm, even if the performed user studies has limitations. Incremental research which reproduces and advances an already published system can strengthen validity of results and ultimately leads to more mature artifacts. As an important support for this, researchers should be able to receive credit beyond citation of publications. Platforms like **Open Science Framework** (**OSF**) or Zenodo allow the assignment of DOIs to artifacts (software, data, or technical descriptions) which are not peer-reviewed articles and these can be referenced in a work that makes use of them. Finally, it should be noted that building on existing work can also mean to follow-up research on own studies, for example to confirm the results of an exploratory experiment through validation with independent metrics (e.g., validating a system for workload classification through a user study measuring efficiency).

Data and Code Sharing

One way of reporting experiments, which seems to be still underused in the community is the sharing of data and code for analysis and experimental paradigms. It is important to further explore the challenges, such as the additional effort it takes to prepare the material with little perceived benefit or recognition. In addition, there may be concerns of increased scrutiny or loss of competitive edge in an academic race for high-impact publications. An "open" culture of publishing material could and should lead to a paradigm change, allowing the publication and joint improvement of research contributions: When an expert in empirical HCI develops an innovative experimental paradigm but an expert in biomedical engineering identifies areas to improve the published data preprocessing pipeline, both should be able to get credit for an updated joint work in contrast to work that is never published.

To move toward increased data sharing in the HCI community, however, there is a need for additional author and reviewer guidance and support. In some venues, particularly the main research conferences, the anonymous review is expected, discouraging authors from sharing external links that may de-anonymize them. However, the article submission systems often do not easily support the submission of data, code, metadata, and so on. An author submitting a article might not know the best way to anonymously share their data for review. Similarly, reviewers need guidance about what level to examine the data, code, and other materials during the review process.

Publishing Expanded Experiment Descriptions

Another recommendation for HCI articles with brain signals is to consider publishing expanded experiment descriptions beyond the conference or journal article. The analysis of the articles shows that no individual article achieved more than 72% of coverage of all eligible model aspects. The missing details limit the ability for other researchers to easily build on the work. For example, a article by Jones, et al. [65] is detailed in its reporting of technical aspects during EEG recording. Many newer articles omit many of these details. However, the experiment from 2003 is also relatively simple compared to modern experiment designs, such as [6], which in turn covers other aspects. The analysis shows that journal publications exhibit on average a better coverage of experiment model aspects compared to conference proceedings. This is likely due to the limitations in space, as brought up in Section 7 by study participants. We, therefore, recommend to make use of the opportunity to publish supplementary material as offered by many conferences and journals. In addition, publication in open data repositories gives the opportunity to share concrete information. Both approaches allow to provide unlimited and more technical information. Alternatively, authors may consider accompanying blog posts or other forms of documentation beyond the publication manuscript itself. Future work needs to study obstacles for making data available [143],

especially such reasons which are specific to brain data; one example is the question of long-term anonymity of brain data, which has implications on the ability to share such data liberally in the context of data protection and privacy laws.

Applying the Experiment Model

One outcome of the experiment model is that researchers can now compare past and future HCI publications against the model to identify strengths and weaknesses in reporting. The model was cross-checked through the analysis with 110 HCI publications to cover a wide range of information that is considered relevant to the HCI community for experiment reporting. It can therefore act as a reference for deciding what information to present in a publication or experiment description. This can help researchers to make their research more accessible and reproducible. We also recommend the consideration of aspects of the model which did not apply to your research, but for which there may be uncertainty if unreported. An example of such a situation is "participants restraints": If a researcher did not restrain participants in any way, it may not occur to them to report this in the article; however, other researchers may wonder if no restraints were applied or whether they were not reported. Furthermore, our database can help researchers identify "role models" for certain types of experiments and certain aspects of experiment reporting. Finally, the experiment model can also be used by reviewers by providing a check list to guide the systematic and objective assessment of experiment reporting.

Further Refinement and Expansion of Experiment Model

We have presented an experiment model as a starting point for further discussion, critique, and extension of the model, as well as identification of the importance of its attributes for different use cases. One source for improvement might be an alternative structure or even a multi-dimensional structure, as the statistical analysis of the articles, showed multiple potential approaches to group the model attributes into categories.

Despite being based on 110 articles, the model cannot cover the full breadth of all relevant articles and future use cases. For example, the database currently only contains one article of multi-user BCI and thus does not reflect the specific aspects of such BCIs, for example how the different users communicate. Multiparty experiments are complex and may require the experiment model to cover additional aspects (e.g., in the roles of the different participants) as these designs become more common. As other novel study designs emerge, the experiment model would need to adapt to represent these new paradigms. Another possible area of refinement is the description of statistical analysis and the discrimination of exploratory and confirmatory aspects, which could be formalized beyond our presented analysis (Section 4.2.9), once an explicit discussion of these aspects becomes more common. Another limitation of our current analysis is that we did not consider articles in domain-specific publication outlets, such as neuroergonomics or learning research, as we wanted to focus on HCI venues. From the identified core domains (see Section 4), we can in the future extend the analysis to publications in these areas.

Finally, the model could also evolve more in the direction of a broader contextualization of research, documenting how researchers discuss their specific experiment in relation to ethical concerns and societal implications [40, 59, 60, 67] revolving around the use of brain data in HCI. Several specialized works present a systematic discussion of these aspects, and list the following characteristics of ethical issues related to BCIs: *personhood*, *stigma*, *autonomy*, *privacy*, *research ethics*, *safety*, *responsibility*, and *justice* [20, 30]. However, they also note that the discussion of ethical aspects is not widespread in empirical articles. This is also reflected in our database: Apart from *informed consent*, which is the most common one and already covered by the experiment model, we selected *autonomy*, *safety,* and *privacy* as the most tangible ones from the list above

and identified the articles which referred to them. When accounting for false positives (e.g., a discussion about "autonomous driving"), we found that 11.8% of all articles referred to safety (for example in reference to safety-critical BCI-applications), 6.3% referred to privacy (often in the context of the data collection), and 5.5% referred to autonomy (for example, in the context of users with disabilities). This count is also generous, because sometimes the concepts are only mentioned briefly in the introduction or discussion.

9 CONCLUSION

In this article, our motivation was to characterize the current state of affairs to facilitate future discussion on reproducibility and reuse of HCI research with brain signals. We studied the diversity of HCI research using brain signals, with regards to domains and applications, modalities, measured mental states and processes, and more. From 110 publications since 1996, we showed the large variety and thus the broad applicability of brain activity measurement to improve or quantify HCI. The analysis also revealed that the studied field is heterogeneous and composed of sub-categories, which use different ways of reporting their experiments. We conclude that these differences may pose a challenge to understand, reproduce, and build on this research.

One result of this work is the creation of an initial experiment model, which acts as a unified superset of recurring aspects of empirical work on using brain signals for HCI purposes. This model and the example attributes can act as a guideline to structure a report on empirical work in this area. It can help to reduce the mental workload and uncertainty for both authors and reviewers by providing structure for publications. Explicating, structuring, and naming different aspects of the model can facilitate a discussion in the community on what and how to report. The empirical analysis has confirmed the general structure of the proposed model but also revealed potential alternatives for future explorations. We saw that while some parts of the model define a minimal standard most articles report on, there are other aspects that are only addressed by parts of the community. A discussion and re-iteration in the scientific community would be an important follow-up from this starting point to create a more mature and comprehensive version of the model. This requires a broadening of perspectives beyond the scope of this article, e.g., through a workshop series at one of the premier HCI conferences. We made the model accessible through a GitHub repository at https://brain-signals-hci.github.io/experiment-model/, through which interested researchers can send pull requests to submit suggestions for improvements.

While there is much room for subsequent research, the presented work is an initial step toward reproducibility and reuse of HCI research with brain signals.

APPENDIX

A DETAILED EXPERIMENT MODEL WITH EXAMPLES

In the tables below, we present the experiment model for HCI research with brain signals, described in Section 5. We describe each category and list the attributes contained within them. For each attribute, we give a definition of the attribute and present an example taken from the surveyed literature. We tried to identify examples which are biased towards a more detailed documentation, but individual examples may still lack certain information. An online version which allows to submit changes is available at https://brain-signals-hci.github.io/experiment-model/.

Table 6. Experiment Model: Part 1

Aspect of Model	Example
Technical Aspects of Recording	
Type of Sensor: For a given brain sensing modality, report the manufacturer and the specification of the sensor chain employed.	*"The EEG was recorded using a NeuroScan system with 32 Ag-AgCl electrodes" (Lee et al. [80])*
Sensor Position: Report where on the scalp electrodes are positioned. For EEG, this is most often done in terms of the 10–20 positioning system or its refinements. For fNIRS, the placement of transmitters and receivers has to be distinguished and the respective distances need to be reported.	*"Electrodes were positioned according to the extended 10-20 system on CPz, POz, Oz, Iz, O1 and O2" (Evain et al. [39])*
Sampling Rate: Report the number of samples recorded per second in Hz.	*"A sampling rate of EEG signals was set as 300 Hz." (Terasawa et al. [132])*
Measurement Quality: For EEG, the threshold for the maximum impedance level (in $k\Omega$) is often reported. For fNIRS, no standardized quality measurement exists, different devices provide different ways of measurement (e.g., photon count).	*"electrode impedance was below 5 KΩ" (Vi et al. [141])*
Reference: Specific to the EEG signal, it is custom to report the electrode to which the recording was referenced.	*"Two electrodes were located at both earlobes as reference and ground." (Terasawa et al. [132])*
Auxiliary signals: The brain sensing modality may not be the only signal, which is captured during the experiment. Often, other sensors, such as eye trackers or heart rate monitors are employed. For these, similar information as for the brain sensing modality may be reported, especially about the specific type of sensor and its placement.	*"Eye positions were measured with an embedded infrared eye-tracking module: aGlass DKI from 7invensun (https://www.7invensun.com)." (Ma et al.[84])*
Synchronization with stimuli and other signals: For analyzing a continous stream of brain signal data, it needs to be synchronized to the events of the experiments (and potentially any other signal sources). This can be done through timestamps, trigger signals, light sensors or other means and the method may be reported to determine the precision of the achieved synchronization.	*"A parallel port connection between recording PC and experimental PC synchronized the EEG recording with the experimental events, such as the sound onset and button press." (Glatz et al. [49])*
Recording Environment: This point reports where and under what conditions the experiment was conducted. Of relevance can be the conditions regarding control of light, sound, electromagnetic fields as well as the positioning of the participant. This attribute is often illustrated through a photo or video of the environment.	*"#Scanners was presented in an intimate 6 person capacity cinema, within a caravan [...]. The space had no windows, low lighting, plush seating, an eight foot projected image, and stereo speakers. Figure 3 [shows a] participant wearing the EEG device, experiencing #Scanners inside the caravan." (Pike et al.[103])*

Table 7. Experiment Model: Part 2

Aspect of Model	Example
Task Description	
Participant Restraints: This attribute relates to any instructions or physical restraints, which were in place during the experiment to avoid artifacts or other undesired effects influencing the signal.	*"the participants were instructed to refrain from excessive movement by keeping arms at rest on the table in a position that allowed them to reach the keyboard without excessive movement." (Crk et al. [31])*
Output devices: Describes through which devices (e.g., computer screen, mobile phone, and so on.) information and material is communicated to the user. As many brain activity patterns are sensitive to the specific characteristics of the stimulation, details of the presentation may be reported.	*"The [...] game stimulus was run on a powerful high-end gaming PC (CPU: Intel® Core™ i7-6850K @ 3.60 GHz; RAM: 32 GB; GPU: NVIDIA Geforce GTX 1080) and displayed on a 27-inch BenQ ZOWIE XL2720 144 hz gaming monitor at a 1920x1080 resolution." (Terkildsen and Makransky,[133])*
Input devices: Describes through which devices (besides the brain signal itself) the user communicates commands and other types of input to the system.	*"HMD-mounted Leap Motion (https://www.leapmotion.com) to track participants' hands." (Škola and Liarokapis [142])*
User Input: Describes which kind of commands and input users can enter into the system at which point of the task. May specify which input devices are used and whether there are any requirements or restrictions to the input.	*"They had to respond to auditory notifications whenever one was presented, with a button press using either their left or right index fingers. Six notifications (i.e., 3 complementary pairs of verbal commands and auditory icons) were pre-assigned to a left index-finger press and the remaining six, to a right index-finger press." (Glatz et al. [49])*
Middleware/Communication: For interactive applications or distributed recording setups, this attribute reports how the different parts communicate to exchange data, triggers, commands, and so on.	*"We wrote a custom Java bridge program to connect the headset to the Android OS and Unity application on the Game tablet. The Java program polled the headset 60 times a second for EEG power spectrum [...] We connected the Calibrate tablet to the Game tablet using WiFi Direct [...]." (Antle et al. [6])*
Framework/Technical platform: What software or development toolkit (in what version) was used as the foundation to implement the task (e.g., PsychoPy, Unity, and so on.)	*"The scene was developed using Unity version 2017.3.0f3, for the representation of hands, the realistically looking hand models "Pepper Hands" from Leap Motion suite were used (visible in Figure 3)." (Škola and Liarokapis [124])*
Task Functionality: Reports what functionality the involved software provides to the user (in the case of a working interactive application) and how it responds to different user input. For experiments which are based on or inspired by established paradigms (e.g., from cognitive psychology), this source may be reported (e.g., in reference to a source such as the Cognitive Atlas [105].	*"the main task for the study is a multi-robot version of the task introduced in [27]. Participants remotely supervised two robots (the blue robot and the red robot) that were exploring different areas of a virtual environment. Participants were told that the two robots had collected information that needed to be transmitted back to the control center. [continues...]" (Solovey et al. [126])*
Architecture: For experiments which involve non-trivial custom software artifacts, this attribute reports the underlying software architecture, informing about structure of and information flow between modules.	

Table 8. Experiment Model: Part 3

Aspect of Model	Example
Stimulus Material: For tasks which involve the repeated presentation of uniform stimuli (e.g., pictures to rate, text prompts to enter, and so on.), this attribute reports the form of these stimuli (e.g., picture size, length, language, and so on.) and their source.	*"To prepare experimental materials, a dataset of notifications from the websites Notification Sounds and Appraw were collected. Seven musically trained raters were recruited to determine the melody complexity of the 40 notifications. [...] (The stimuli can be downloaded at https://goo.gl/SnZrzG)." (Cherng et al. [25])*
Visualization provided?: This attribute reports visually (through screen shots or video) the task as shown to the user. If the task has multiple distinct parts, all of them may be visualized. If the task is not in English, the visualization may be accompanied by a translation.	
Timing: Reports on if and how the task is (partially) paced by an internal clock, for example, for controlling the duration for stimulus presentation or the time time for responding to a prompt.	*"Each trial began with a black screen for 3s, followed by a fixation dot in the center of the screen for 200ms. After that, the screen remains clear for 200ms before one of four stimuli was displayed for 300ms." (Vi et al. [140])*
Code for task provided?: Reports on whether the task is provided in source code or an executable file and under what licence. If custom hardware is involved, this could also include a blueprint or a circuit diagram.	
Participants	
Recruitment strategy: How where study participants recruited, e.g., through social media, in class, and so on.?	*"A snowball procedure was used to gather the sample of study participants. The study was advertised via university courses, email and social media." (Johnson et al. [63])*
Incentives: What compensation (if any) was offered to study participants, e.g., money, class credit, and so on.? What were the criteria for being eligible for the compensation?	*"Participants received monetary compensation for their participation (10 Euro)." (Putze et al. [112])*
Age: How old are the participants (mean and standard deviation)?	*"mean age 24.53 (SD: 3.00)" (Frey et al. [44])*
Gender: With what gender do participants identify (relative frequencies)?	*"2 females and 9 males" (Ma et al. [84])*
Occupation: What is the profession or—in case of students—the field of study of the participants?	*"Data were collected from 34 computer science undergraduates at the first two authors' institution" (Crk et al. [31])*

Table 9. Experiment Model: Part 4

Aspect of Model	Example
Inclusion or exclusion criteria: Where there rules on which participants were eligible to take part in the experiment and what were these criteria (e.g., handedness, disabilities, caffeine consumption, and so on.)?	*"Each of the individuals was enrolled in at least one computer science course" (Crk et al. [31])*
Approval of ethics committee: Was the study approved by an ethics committee? If so, by which one?	*"the experimental protocol was approved by the University Research Ethics Committee prior to data collection." (Burns and Fairclough, [19])*
Experiment Flow	
Experiment Structure: Order of different segments of the experiment, such as instruction, sensor placement, training, debriefing, and so on. For "in-the-wild" experiments, which do not follow a fixed, predefined pattern, this attribute may report the boundary conditions and which parts of the experiment could be.	*"•5-10min of explanation (slideshow with a detailed description of the interface), •5min to set up the hardware, •For each BCI paradigm: 10min of calibration, familiarization •10min to play the game •5min of rest •Questionnaires" (Kosmyna et al. [69])*
Instructions: Instructions given to the participants regarding purpose of experiment, experiment setup, task operations, restrictions, safety considerations, and so on. If possible, the written instruction documents may be provided.	*"Participants were instructed to focus visual attention on a target symbol, whilst silently counting the number of times the target character flashed." (Obeidat et al. [98])*
Training procedure: This attribute reports about how the participants familiarized themselves with the task. it may report the duration of training, specific training conditions compared to the main experiment, and specific training instructions.	*"All participants first undertook a training task where they played 15 easy and 15 hard pieces on the piano. Each piece was 30 seconds long with a 30-second rest between each piece." (Yuksel et al. [152])*
Trial ordering: For any experiment that is organized in trials, this attribute reports how the ordering of the blocks was derived. It may report whether and how ordering differed between participants.	*"The order of conditions was counter-balanced across subjects and participants wore the fNIRS device during both trials so they did not notice a difference between the two conditions." (Afergan et al. [1])*
Repetitions, blocks & breaks: This attribute reports the larger structure of the experiments, such as blocks of trials and their duration, ordering as well as pauses between blocks. If applicable also reports stopping criteria, if these are individual.	*"Each session comprised four testing blocks: two using club ambient noise and two using city street ambient noise as standard stimuli. Between any two consecutive blocks, subjects had three minutes to rest." (Lee et al. [79])*
Pre-study screenings: Reports any procedures prior to the experiment that determine the eligibility of a participant for the study, their assignments to an experiment group or other aspects of the experiment process.	*"None of the subjects had any history of brain disease, drug use, or hearing problems. None had any musical expertise." (Lee et al. [79])*

Table 10. Experiment Model: Part 5

Aspect of Model	Example
Questionnaires: Reports which questionnaires were administered before, during and after the experiment. May give a reference to a published questionnaire or list the items of a custom one. May also report when and how often questionnaires were administered and through what means (article or computer-based).	*"At the end of the evaluation, end-users were asked to complete the NASA-TLX, the eQUEST 2.0, and a customized usability questionnaire. [...] After each session was complete, the therapist station would automatically open up on the laptop and ask the user to answer, How satisfied were you with the BCI session? (with 10 being very satisfied and 0 is not satisfied) [continues ...]" (Miralles et al. [90])*
Data Processing	
Derivation of labels: Outside neurofeedback applications the recorded brain signal data is distributed between multiple groups or assigned a continuous value. This attribute may report how the label is derived from the collected data (e.g., defined by the experiment structure, by questionnaire responses, or external ratings).	*"We considered the mean of the three NASA-TLX parameters (effort, mental demand and frustration) to evaluate the overall mental workload. The average score was thresholded at the mean value of 2 (since the used scale was 0–4) to quantize or characterize a parameter block as inducing low/high workload." (Bilalpur et al. [14])*
Data transformation: This attribute refers to all processing steps which transform raw data while keeping it in the original time-domain representation. Examples of such transformation steps are: re-referencing, baseline normalization, downsampling, and so on.	*"the common average was subtracted from all EEG channels." (Lampe et al. [76])*
Filtering: This attribute reports any filtering of the data. This may include the type of filter applied as well as necessary parameters, such as the filter order.	*"EEG data was first low-pass filtered with a cut-off frequency of* $50hz$ *and high-pass filtered with a cutoff frequency of* $0.16hz$*, both using a third-order butterworth filter" (Rodrigue et al. [118])*
Windowing: This attribute reports how segments of data are aligned (e.g., locked to an event in the experiment), how long they are and with which window function they are extracted.	*"The data was then segmented into 1.5-second epochs, overlapping each previous epoch by* 50%*" (Rodrigue et al. [118])*
Artifact cleaning: Reports through which algorithms (beyond filtering) artifacts were removed and which artifacts are targeted.	*"Independent Component Analysis (ICA) is applied [...] The components are first filtered using a band-pass filter with cut off frequencies 1 - 6 Hz. Choosing the component with the highest energy [and] applying a high-pass filter with a cut off frequency of 20 Hz." (Jarvis et al. [61])*
Hyperparameter optimization: For machine learning models, this attribute reports how the hyperparameters of the model were chosen (e.g., through grid search) and which hyperparameters where chosen in the final model. This also includes other parameters of the processing pipeline which are optimized (e.g., in preprocessing).	*"A grid search was performed to optimize sigma for all participants, the remaining parameters were left as default." (Rodrigue et al. [118])*

Table 11. Experiment Model: Part 6

Aspect of Model	Example
Outlier handling: Reports any methods for excluding certain samples, windows, or sessions based on the contained data or other external factors.	*"Any rest or trial period with 20 percent or higher error rate is considered noisy and can be excluded from the analysis" (Crk et al. [31])*
Feature extraction: This attribute reports on how a feature vector for classification or regression is calculated from the preprocessed data.	*"[W]e partitioned each data window into smaller segments of 50 ms length. We then used the signal mean of the segment, calculated on the band-pass filtered signal, with cutoff frequencies at 4 and 13 Hz (i.e. θ- and α-bands)." (Putze et al. [112])*
Feature selection: This attribute reports on procedures to reduce the number of features automatically.	*"we performed a feature selection using the Fisher ratio as selection criterion. The number k of selected features [...] was a tuning parameter in the range between 5 and 50." (Putze et al. [112])*
Learning model: This attribute reports on the specific machine learning model that is employed (if any) to perform classification or regression.	*"the Neural Network Toolbox of MATLAB was used to create an artificial neural network (ANN) with 198 inputs, 20 hidden neurons and 4 outputs. The patternnetfunction, which creates a feed-forward neural network, was used. [...]" (Lampe et al. [76])*
Evaluation procedure: For machine learning models, this attribute reports how they were evaluated to assess their performance. This involves the exact metric used for assessment as well as the approach to (sometimes repeatedly) determine test and training datasets.	*"To assess the classifiers' performance on the calibration data, we used 4-fold cross-validation (CV). [...] The performance was measured using the area under the receiver-operating characteristic curve (AUROCC)." (Frey et al. [44])*
Processing code provided? Reports if the code for processing the brain signal data is released with the paper or in a separate repository. If the code cannot be provided, as a substitute it is possible to report the employed frameworks (e.g., EEGLAB).	*"The full classification pipeline is implemented in Python. For EEG processing, we use the MNE toolbox [17]. For machine learning and evaluation algorithms, we use scikit [28] and custom routines build on numpy and scipy." (Putze et al. [112])*
Brain Signal Integration	
Brain Input effect: This attribute describes how the output of the brain input processing influences the design, the behavior, or the content of the application or experimental paradigm.	*"when the system was confident that the user was in a state of low or high workload, one UAV would be added or removed, respectively. After a UAV was added or removed, there was a 20 second period where no more vehicles were added or removed." (Afergan et al. [1])*
Type of integration: This attribute describes the algorithmic implementation of the brain signal integration, i.e., whether an explicit conditional statement, an Influence Diagram, a state graph, or a different way of behavior modeling was used.	*"[The self-correction algorithm] inspects the probability distribution [...] and picks the now highest scoring class [...]. [W]e only used the second best class if its re-normalized confidence [...] is above a certain threshold T [...]. Otherwise, the user was asked to repeat the input." (Putze et al. [107])*

ACKNOWLEDGMENTS

The authors would like to thank Jayesh Dubey for his work.

REFERENCES

[1] Daniel Afergan, Evan M. Peck, Erin T. Solovey, Andrew Jenkins, Samuel W. Hincks, Eli T. Brown, Remco Chang, and Robert J. K. Jacob. 2014. Dynamic difficulty using brain metrics of workload. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3797–3806. DOI: https://doi.org/10.1145/2556288.2557230

[2] Daniel Afergan, Tomoki Shibata, Samuel W. Hincks, Evan M. Peck, Beste F. Yuksel, Remco Chang, and Robert J.K. Jacob. 2014. Brain-based target expansion. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 583–593. DOI: https://doi.org/10.1145/2642918.2647414

[3] Akshay Aggarwal, Gerrit Niezen, and Harold Thimbleby. 2014. User experience evaluation through the brain's electrical activity. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. ACM, New York, NY, 491–500. DOI: https://doi.org/10.1145/2639189.2639236

[4] Jennifer Allanson and Stephen H Fairclough. 2004. A research agenda for physiological computing. *Interacting with Computers* 16, 5 (2004), 857–878.

[5] Bonnie Brinton Anderson, C. Brock Kirwan, Jeffrey L. Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 2883–2892. DOI: https://doi.org/10.1145/2702123.2702322

[6] Alissa N. Antle, Leslie Chesick, and Elgin-Skye Mclaren. 2018. Opening up the design space of neurofeedback brain–computer interfaces for children. *ACM Transactions on Computer-Human Interaction* 24, 6 (2018), 33 pages. DOI: https://doi.org/10.1145/3131607

[7] Gabor Aranyi, Fred Charles, and Marc Cavazza. 2015. Anger-based BCI using fNIRS neurofeedback. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, New York, NY, 511–521. DOI: https://doi.org/10.1145/2807442.2807447

[8] P Aricò, G Borghini, G Di Flumeri, N Sciaraffa, and F Babiloni. 2018. Passive BCI beyond the lab: Current trends and future directions. *Physiological Measurement* 39, 8 (2018), 08TR02.

[9] Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. 2021. A critique of electrodermal activity practices at CHI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[10] Panagiotis D Bamidis, Christos Papadelis, Chrysoula Kourtidou-Papadeli, Costas Pappas, and Ana B. Vivas. 2004. Affective computing in the era of contemporary neurophysiology and health informatics. *Interacting with Computers* 16, 4 (2004), 715–721.

[11] Danushka Bandara, Senem Velipasalar, Sarah Bratt, and Leanne Hirshfield. 2018. Building predictive models of emotion with functional near-infrared spectroscopy. *International Journal of Human-Computer Studies* 110 (2018), 75–85. DOI: https://doi.org/10.1016/j.ijhcs.2017.10.001

[12] Oswald Barral, Ilkka Kosunen, and Giulio Jacucci. 2017. 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* 24, 6 (2017), 29 pages. DOI: https://doi.org/10.1145/3157730

[13] Joanna Bergström, Tor-Salve Dalsgaard, Jason Alexander, and Kasper Hornbæk. 2021. How to evaluate object selection and manipulation in VR? Guidelines from 20 years of studies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–20.

[14] Maneesh Bilalpur, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. 2018. EEG-based evaluation of cognitive workload induced by acoustic parameters for data sonification. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, 315–323. DOI: https://doi.org/10.1145/3242969.3243016

[15] Mark Boyer, Mary L Cummings, Lee B Spence, and Erin T Solovey. 2015. Investigating mental workload changes in a long duration supervisory control task. *Interacting with Computers* 27, 5 (2015), 512–520.

[16] Anne-Marie Brouwer, Thorsten O Zander, Jan BF Van Erp, Johannes E Korteling, and Adelbert W Bronkhorst. 2015. Using neurophysiological signals that reflect cognitive or affective state: Six recommendations to avoid common pitfalls. *Frontiers in Neuroscience* 9 (2015), 136.

[17] Clemens Brunner, Niels Birbaumer, Benjamin Blankertz, Christoph Guger, Andrea Kübler, Donatella Mattia, José del R. Millán, Felip Miralles, Anton Nijholt, Eloy Opisso, Nick Ramsey, Patric Salomon, and Gernot R. Müller-Putz. 2015. BNCI Horizon 2020: Towards a roadmap for the BCI community. *Brain-computer Interfaces* 2, 1 (2015), 1–10.

[18] Derrick Matthew Buchanan, Jeremy Grant, and Amedeo D'Angiulli. 2019. Commercial wireless versus standard stationary EEG systems for personalized emotional brain-computer interfaces: A preliminary reliability check. *Neuroscience Research Notes* 2, 1 (2019), 7–15. DOI: https://doi.org/10.31117/neuroscirn.v2i1.21

[19] Christopher G Burns and Stephen H Fairclough. 2015. Use of auditory event-related potentials to measure immersion during a computer game. *International Journal of Human-Computer Studies* 73 (2015), 107–114. DOI: https://doi.org/10.1016/j.ijhcs.2014.09.002

[20] Sasha Burwell, Matthew Sample, and Eric Racine. 2017. Ethical aspects of brain computer interfaces: A scoping review. *BMC Medical Ethics* 18, 1 (2017), 1–11.

[21] Raymundo Cassani, Hubert Banville, and Tiago H. Falk. 2015. MuLES: An open source EEG acquisition and streaming server for quick and simple prototyping and recording. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*. ACM, New York, NY, 9–12. DOI: https://doi.org/10.1145/2732158.2732193

[22] Christopher D Chambers, Eva Feredoes, Suresh Daniel Muthukumaraswamy, and Peter Etchells. 2014. Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience* 1, 1 (2014), 4–17.

[23] Guillaume Chanel, Joep JM Kierkels, Mohammad Soleymani, and Thierry Pun. 2009. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies* 67, 8 (2009), 607–627.

[24] Guillaume Chanel and Christian Mühl. 2015. Connecting brains and bodies: Applying physiological computing to support social interaction. *Interacting with Computers* 27, 5 (2015), 534–550.

[25] Fu-Yin Cherng, Yi-Chen Lee, Jung-Tai King, and Wen-Chieh Lin. 2019. Measuring the influences of musical parameters on cognitive and behavioral responses to audio notifications using EEG and large-scale online studies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 12 pages. DOI: https://doi.org/10.1145/3290605.3300639

[26] Fu-Yin Cherng, Wen-Chieh Lin, Jung-Tai King, and Yi-Chen Lee. 2016. An EEG-based approach for evaluating graphic icons from the perspective of semantic distance. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 4378–4389. DOI: https://doi.org/10.1145/2858036.2858133

[27] S. Mealla Cincuegrani, S. Jordà, and A. Väljamäe. 2016. Physiopucks: Increasing user motivation by combining tangible and implicit physiological interaction. *ACM Transactions on Computer-Human Interaction* 23, 1 (2016), 22 pages. DOI: https://doi.org/10.1145/2838732

[28] Miriam Clemente, Beatriz Rey, Aina Rodríguez-Pujadas, Alfonso Barros-Loscertales, Rosa M Baños, Cristina Botella, Mariano Alcañiz, and César Ávila. 2014. An fMRI study to analyze neural correlates of presence during virtual reality experiences. *Interacting with Computers* 26, 3 (2014), 269–284.

[29] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK no more: On the preregistration of CHI experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 141:1–141:12. DOI: https://doi.org/10.1145/3173574.3173715

[30] Allen Coin, Megan Mulder, and Veljko Dubljević. 2020. Ethical aspects of BCI technology: What is the state of the art? *Philosophies* 5, 4 (2020), 31.

[31] Igor Crk, Timothy Kluthe, and Andreas Stefik. 2015. Understanding programming expertise: An empirical study of phasic brain wave changes. *ACM Transactions on Computer-Human Interaction* 23, 1 (2015), 29 pages. DOI: https://doi.org/10.1145/2829945

[32] Edward Cutrell and Desney Tan. 2008. BCI for passive input in HCI. In *Proceedings of CHI 2008 Workshop on Brain-Computer Interfaces for HCI and Games*. 1–3.

[33] Tiziano D'albis, Rossella Blatt, Roberto Tedesco, Licia Sbattella, and Matteo Matteucci. 2012. A predictive speller controlled by a brain-computer interface based on motor imagery. *ACM Transactions on Computer-Human Interaction* 19, 3 (2012), 25 pages. DOI: https://doi.org/10.1145/2362364.2362368

[34] Yi Ding, Brandon Huynh, Aiwen Xu, Tom Bullock, Hubert Cecotti, Matthew Turk, Barry Giesbrecht, and Tobias Höllerer. 2019. Multimodal classification of EEG during physical activity. In *Proceedings of the 2019 International Conference on Multimodal Interaction*. 185–194.

[35] Eamon Doherty, Gilbert Cockton, Chris Bloor, and Dennis Benigno. 2001. Improving the performance of the cyberlink mental interface with "Yes/no program". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 69–76. DOI: https://doi.org/10.1145/365024.365038

[36] Eamon P Doherty, Gilbert Cockton, Chris Bloor, Joann Rizzo, Bruce Blondina, and Bruce Davis. 2002. Yes/no or maybe–further evaluation of an interface for brain-injured individuals. *Interacting with Computers* 14, 4 (2002), 341–358.

[37] Florian Echtler and Maximilian Häußler. 2018. Open source, open science, and the replication crisis in HCI. In *Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–8. DOI: https://doi.org/10.1145/3170427.3188395

[38] Ernest Edmonds, Dave Everitt, Michael Macaulay, and Greg Turner. 2004. On physiological computing with an application in interactive art. *Interacting with Computers* 16, 5 (2004), 897–915.

[39] Andéol Evain, Ferran Argelaguet, Nicolas Roussel, Géry Casiez, and Anatole Lécuyer. 2017. Can I think of something else when using a BCI? Cognitive demand of an SSVEP-based BCI. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5120–5125. DOI: https://doi.org/10.1145/3025453.3026037

[40] Stephen Fairclough. 2014. Physiological data must remain confidential. *Nature News* 505, 7483 (2014), 263.
[41] Stephen H. Fairclough. 2009. Fundamentals of physiological computing. *Interacting with Computers* 21, 1–2 (2009), 133–145.
[42] Adam R. Ferguson, Jessica L. Nielson, Melissa H. Cragin, Anita E. Bandrowski, and Maryann E. Martone. 2014. Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience* 17, 11 (2014), 1442–1447.
[43] Stéphanie Fleck and Martin Hachet. 2016. Making tangible the intangible: Hybridization of the real and the virtual to enhance learning of abstract phenomena. *Frontiers in ICT* 3 (2016), 30 pages. DOI: https://doi.org/10.3389/fict.2016.00030
[44] Jérémy Frey, Maxime Daniel, Julien Castet, Martin Hachet, and Fabien Lotte. 2016. Framework for electroencephalography-based evaluation of user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 2283–2294. DOI: https://doi.org/10.1145/2858036.2858525
[45] Jérémy Frey, Renaud Gervais, Stéphanie Fleck, Fabien Lotte, and Martin Hachet. 2014. Teegi: Tangible EEG interface. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 301–308. DOI: https://doi.org/10.1145/2642918.2647368
[46] Doron Friedman, Anne-Marie Brouwer, and Anton Nijholt. 2017. BCIforReal: An application-oriented approach to BCI out of the laboratory. In *Proceedings of the 22nd international conference on intelligent user interfaces companion*. 5–7.
[47] Lukas Gehrke, Sezen Akman, Pedro Lopes, Albert Chen, Avinash Kumar Singh, Hsiang-Ting Chen, Chin-Teng Lin, and Klaus Gramann. 2019. Detecting visuo-haptic mismatches in virtual reality using the prediction error negativity of event-related brain potentials. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 11 pages. DOI: https://doi.org/10.1145/3290605.3300657
[48] Ioana Ghergulescu and Cristina Hava Muntean. 2014. A novel sensor-based methodology for learner's motivation analysis in game-based learning. *Interacting with Computers* 26, 4 (2014), 305–320.
[49] Christiane Glatz, Stas S. Krupenia, Heinrich H. Bülthoff, and Lewis L. Chuang. 2018. Use the right sound for the right job: Verbal commands and auditory icons for a task-management system favor different information processes in the brain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 13 pages. DOI: https://doi.org/10.1145/3173574.3174046
[50] David Grimes, Desney S. Tan, Scott E. Hudson, Pradeep Shenoy, and Rajesh P.N. Rao. 2008. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 835–844. DOI: https://doi.org/10.1145/1357054.1357187
[51] Hayrettin Gürkök, Gido Hakvoort, and Mannes Poel. 2011. Modality switching and performance in a thought and speech controlled computer game. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, New York, NY, 41–48. DOI: https://doi.org/10.1145/2070481.2070491
[52] Nils Hachmeister, Hannes Riechmann, Helge Ritter, and Andrea Finke. 2011. An approach towards human-robot-human interaction using a hybrid brain-computer interface. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, New York, NY, 49–52. DOI: https://doi.org/10.1145/2070481.2070492
[53] Peter Halling, Paul F. Fitzpatrick, Frank M. Raushel, Johann Rohwer, Santiago Schnell, Ulrike Wittig, Roland Wohlgemuth, and Carsten Kettner. [n.d.]. An empirical analysis of enzyme function reporting for experimental reproducibility: Missing/incomplete information in published papers. 242 ([n. d.]), 22–27. DOI: https://doi.org/10.1016/j.bpc.2018.08.004
[54] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. 2017. EngageMeter: A system for implicit audience engagement sensing using electroencephalography. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5114–5119. DOI: https://doi.org/10.1145/3025453.3025669
[55] Leanne M. Hirshfield, Rebecca Gulotta, Stuart Hirshfield, Sam Hincks, Matthew Russell, Rachel Ward, Tom Williams, and Robert Jacob. 2011. This is your brain on interfaces: Enhancing usability testing with functional near-infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 373–382. DOI: https://doi.org/10.1145/1978942.1978996
[56] Leanne M. Hirshfield, Erin Treacy Solovey, Audrey Girouard, James Kebinger, Robert J.K. Jacob, Angelo Sassaroli, and Sergio Fantini. 2009. Brain measurement for usability testing and adaptive interfaces: An example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 2185–2194. DOI: https://doi.org/10.1145/1518701.1519035
[57] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3523–3532. DOI: https://doi.org/10.1145/2556288.2557004
[58] Jin Huang, Chun Yu, Yuntao Wang, Yuhang Zhao, Siqi Liu, Chou Mo, Jie Liu, Lie Zhang, and Yuanchun Shi. 2014. FOCUS: Enhancing children's engagement in reading by using contextual BCI training sessions. In *Proceedings of*

the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, NY, 1905–1908. DOI: https://doi.org/10.1145/2556288.2557339

[59] Marcello Ienca and Pim Haselager. 2016. Hacking the brain: Brain–computer interfacing technology and the ethics of neurosecurity. *Ethics and Information Technology* 18, 2 (2016), 117–129.

[60] Marcello Ienca, Pim Haselager, and Ezekiel J Emanuel. 2018. Brain leaks and consumer neurotechnology. *Nature Biotechnology* 36, 9 (2018), 805–810.

[61] Jan Jarvis, Felix Putze, Dominic Heger, and Tanja Schultz. 2011. Multimodal person independent recognition of workload related biosignal patterns. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, New York, NY, 205–208. DOI: https://doi.org/10.1145/2070481.2070516

[62] Camille Jeunet, Stefan Debener, Fabien Lotte, Jeremie Mattout, Reinhold Scherer, and Catharina Zich. 2018. Mind the traps! Design guidelines for rigorous BCI experiments. In *Proceedings of the Brain–Computer Interfaces Handbook: Technological and Theoretical Advances*, Chang S. Nam, Anton Nijholt, and Fabien Lotte (Eds.). CRC Press, 613–634.

[63] Daniel Johnson, Peta Wyeth, Madison Clark, and Christopher Watling. 2015. Cooperative game play with avatars and agents: Differences in brain activity and the experience of play. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3721–3730. DOI: https://doi.org/10.1145/2702123.2702468

[64] Andrew R. Jones, Michael Miller, Ruedi Aebersold, Rolf Apweiler, Catherine A. Ball, Alvis Brazma, James DeGreef, Nigel Hardy, Henning Hermjakob, Simon J. Hubbard, Peter Hussey, Mark Igra, Helen Jenkins, Randall K. Julian, Kent Laursen, Stephen G. Oliver, Norman W. Paton, Susanna-Assunta Sansone, Ugis Sarkans, Christian J. Stoeckert, Chris F. Taylor, Patricia L. Whetzel, Joseph A. White, Paul Spellman, and Angel Pizarro. 2007. The functional genomics experiment model (FuGE): An extensible framework for standards in functional genomics. *Nature Biotechnology* 25, 10 (2007), 1127–1133. DOI: https://doi.org/10.1038/nbt1347

[65] Keith S. Jones, Matthew Middendorf, Grant R. McMillan, Gloria Calhoun, and Joel Warm. 2003. Comparing mouse and steady-state visual evoked response-based control. *Interacting with Computers* 15, 4 (2003), 603–621.

[66] Silvia Erika Kober and Christa Neuper. 2012. Using auditory event-related EEG potentials to assess presence in virtual reality. *International Journal of Human-Computer Studies* 70, 9 (2012), 577–587.

[67] Johannes Kögel, Jennifer R Schmid, Ralf J Jox, and Orsolya Friedrich. 2019. Using brain-computer interfaces: A scoping review of studies employing social research methods. *BMC Medical Ethics* 20, 1 (2019), 1–17.

[68] Nataliya Kosmyna and Anatole Lécuyer. 2019. A conceptual space for EEG-based brain-computer interfaces. *PloS One* 14, 1 (2019), e0210145.

[69] Nataliya Kosmyna, Franck Tarpin-Bernard, and Bertrand Rivet. 2015. Adding human learning in brain–computer interfaces (BCIs): Towards a practical control modality. *ACM Transactions on Computer-Human Interaction* 22, 3(2015), 37 pages. DOI: https://doi.org/10.1145/2723162

[70] Nataliya Kosmyna, Franck Tarpin-Bernard, and Bertrand Rivet. 2015. Conceptual priming for in-game BCI training. *ACM Transactions on Computer-Human Interaction* 22, 5 (2015), 25 pages. DOI: https://doi.org/10.1145/2808228

[71] Makrina Viola Kosti, Kostas Georgiadis, Dimitrios A. Adamos, Nikos Laskaris, Diomidis Spinellis, and Lefteris Angelis. 2018. Towards an affordable brain computer interface for the assessment of programmers' mental workload. *International Journal of Human-Computer Studies* 115 (2018), 52–66. DOI: https://doi.org/10.1016/j.ijhcs.2018.03.002

[72] Roman Krepki, Gabriel Curio, Benjamin Blankertz, and Klaus-Robert Müller. 2007. Berlin brain–computer interface—The HCI communication channel for discovery. *International Journal of Human-Computer Studies* 65, 5 (2007), 460–477.

[73] Laurens R. Krol, Sarah-Christin Freytag, and Thorsten O. Zander. 2017. Meyendtris: A hands-free, multimodal tetris clone using eye tracking and passive BCI for intuitive neuroadaptive gaming. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, 433–437. DOI: https://doi.org/10.1145/3136755.3136805

[74] Laurens R. Krol and Thorsten O. Zander. 2017. Passive BCI-based neuroadaptive systems. In *Proceedings of the GBCIC*.

[75] Dean J. Krusienski, Eric W. Sellers, Dennis J. McFarland, Theresa M. Vaughan, and Jonathan R. Wolpaw. 2008. Toward enhanced P300 speller performance. *Journal of Neuroscience Methods* 167, 1 (2008), 15–21.

[76] Thomas Lampe, Lukas D. J. Fiederer, Martin Voelker, Alexander Knorr, Martin Riedmiller, and Tonio Ball. 2014. A brain-computer interface for high-level remote control of an autonomous, reinforcement-learning-based robotic system for reaching and grasping. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 83–88. DOI: https://doi.org/10.1145/2557500.2557533

[77] Johnny Chung Lee and Desney S. Tan. 2006. Using a low-cost electroencephalograph for task classification in HCI research. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 81–90. DOI: https://doi.org/10.1145/1166253.1166268

[78] Yi-Chen Lee, Fu-Yin Cherng, Jung-Tai King, and Wen-Chieh Lin. 2019. To repeat or not to repeat? Redesigning repeating auditory alarms based on EEG analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY 10 pages. DOI: https://doi.org/10.1145/3290605.3300743

[79] Yi-Chieh Lee, Wen-Chieh Lin, Jung-Tai King, Li-Wei Ko, Yu-Ting Huang, and Fu-Yin Cherng. 2014. An EEG-based approach for evaluating audio notifications under ambient sounds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3817–3826. DOI: https://doi.org/10.1145/2556288.2557076

[80] Yi-Chieh Lee, Wen-Chieh Lin, Jung-Tai King, Li-Wei Ko, Yu-Ting Huang, and Fu-Yin Cherng. 2014. An EEG-based approach for evaluating audio notifications under ambient sounds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3817–3826. DOI: https://doi.org/10.1145/2556288.2557076

[81] Fannie Liu, Laura Dabbish, and Geoff Kaufman. 2017. Can biosignals be expressive? How visualizations affect impression formation from shared brain activity. In *Proceedings of the ACM Transactions on Computer-Human Interaction* , (2017), 21 pages. DOI: https://doi.org/10.1145/3134706

[82] Fabien Lotte, Josef Faller, Christoph Guger, Yann Renard, Gert Pfurtscheller, Anatole Lécuyer, and Robert Leeb. 2012. Combining BCI with virtual reality: Towards new applications and improved BCI. In *Proceedings of the Towards Practical Brain-Computer Interfaces*. Springer, 197–220.

[83] Kristiyan Lukanov, Horia A. Maior, and Max L. Wilson. 2016. Using fNIRS in usability testing: Understanding the effect of web form layout on mental workload. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 4011–4016. DOI: https://doi.org/10.1145/2858036.2858236

[84] Xinyao Ma, Zhaolin Yao, Yijun Wang, Weihua Pei, and Hongda Chen. 2018. Combining brain-computer interface and eye tracking for high-speed text entry in virtual reality. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. ACM, New York, NY, 263–267. DOI: https://doi.org/10.1145/3172944.3172988

[85] Horia A. Maior, Matthew Pike, Sarah Sharples, and Max L. Wilson. 2015. Examining the reliability of using fNIRS in realistic HCI settings for spatial and verbal tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3039–3042. DOI: https://doi.org/10.1145/2702123.2702315

[86] Rytis Maskeliunas, Robertas Damasevicius, Ignas Martisius, and Mindaugas Vasiljevas. 2016. Consumer-grade EEG devices: Are they usable for control tasks? *PeerJ* 4 (2016), e1746. DOI: https://doi.org/10.7717/peerj.1746

[87] Nick Merrill and John Chuang. 2018. From scanning brains to reading minds: Talking to engineers about brain-computer interface. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 11 pages. DOI: https://doi.org/10.1145/3173574.3173897

[88] Vojkan Mihajlović. 2019. EEG spectra vs recurrence features in understanding cognitive effort. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 160–165.

[89] Michael P. Milham, R. Cameron Craddock, Jake J. Son, Michael Fleischmann, Jon Clucas, Helen Xu, Bonhwang Koo, Anirudh Krishnakumar, Bharat B. Biswal, F. Xavier Castellanos, et al. 2018. Assessment of the impact of shared brain imaging data on the scientific literature. *Nature Communications* 9, 1 (2018), 1–7.

[90] Felip Miralles, Eloisa Vargiu, Xavier Rafael-Palou, Marc Solà, Stefan Dauwalder, Christoph Guger, Christoph Hintermüller, Arnau Espinosa, Hannah Lowish, Suzanne Martin, Elaine Armstrong, and Jean Daly. 2015. Brain–computer interfaces on track to home: Results of the evaluation at disabled end-users' homes and lessons learnt. *Frontiers in ICT* 2 (2015), 25. DOI: https://doi.org/10.3389/fict.2015.00025

[91] Evan Morgan, Hatice Gunes, and Nick Bryan-Kinns. 2015. Using affective and behavioural sensors to explore aspects of collaborative music making. *International Journal of Human-Computer Studies* 82 (2015), 31–47.

[92] Maryam Mustafa, Stefan Guthe, Jan-Philipp Tauscher, Michael Goesele, and Marcus Magnor. 2017. How human am I? EEG-based evaluation of virtual characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5098–5108. DOI: https://doi.org/10.1145/3025453.3026043

[93] Maryam Mustafa, Lea Lindemann, and Marcus Magnor. 2012. EEG analysis of implicit human visual perception. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 513–516. DOI: https://doi.org/10.1145/2207676.2207746

[94] Andrew Myrden and Tom Chau. 2017. A passive EEG-BCI for single-trial detection of changes in mental state. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 4 (2017), 345–356.

[95] Femke Nijboer, Bram Van De Laar, Steven Gerritsen, Anton Nijholt, and Mannes Poel. 2015. Usability of three electroencephalogram headsets for brain–computer interfaces: A within subject comparison. *Interacting with Computers* 27, 5 (2015), 500–511.

[96] Domen Novak, Benjamin Beyeler, Ximena Omlin, and Robert Riener. 2015. Workload estimation in physical human–robot interaction using physiological measurements. *Interacting with Computers* 27, 6 (2015), 616–629.

[97] Domen Novak, Matjaž Mihelj, and Marko Munih. 2012. A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. *Interacting with Computers* 24, 3 (2012), 154–172.

[98] Qasem Obeidat, Tom Campbell, and Jun Kong. 2013. The Zigzag paradigm: A new p300-based brain computer interface. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. ACM, New York, NY, 205–212. DOI: https://doi.org/10.1145/2522848.2522880

[99] Kenton O'Hara, Abigail Sellen, and Richard Harper. 2011. Embodiment in brain-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 353–362. DOI: https://doi.org/10.1145/1978942.1978994

[100] Martin J. O'Connor, Marcos Martínez-Romero, Attila L. Egyedi, Debra Willrett, John Graybeal, and Mark A. Musen. 2016. An open repository model for acquiring knowledge about scientific experiments. In *Proceedings of the European Knowledge Acquisition Workshop*. Springer, 762–777.

[101] Jaeyoung Park, Kee-Eung Kim, and Sungho Jo. 2010. A POMDP approach to P300-based brain-computer interfaces. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 1–10. DOI: https://doi.org/10.1145/1719970.1719972

[102] Evan M. M. Peck, Beste F. Yuksel, Alvitta Ottley, Robert J. K. Jacob, and Remco Chang. 2013. Using fNIRS brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 473–482. DOI: https://doi.org/10.1145/2470654.2470723

[103] Matthew Pike, Richard Ramchurn, Steve Benford, and Max L. Wilson. 2016. #Scanners: Exploring the control of adaptive films using brain-computer interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5385–5396. DOI: https://doi.org/10.1145/2858036.2858276

[104] Matthew F. Pike, Horia A. Maior, Martin Porcheron, Sarah C. Sharples, and Max L. Wilson. 2014. Measuring the effect of think aloud protocols on workload using fNIRS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3807–3816. DOI: https://doi.org/10.1145/2556288.2556974

[105] Russell A. Poldrack, Aniket Kittur, Donald Kalar, Eric Miller, Christian Seppa, Yolanda Gil, D. Stott Parker, Fred W. Sabb, and Robert M. Bilder. 2011. The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics* 5 (2011), 17.

[106] Riccardo Poli, Caterina Cinel, Ana Matran-Fernandez, Francisco Sepulveda, and Adrian Stoica. 2013. Towards cooperative brain-computer interfaces for space navigation. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. ACM, New York, NY, 149–160. DOI: https://doi.org/10.1145/2449396.2449417

[107] Felix Putze, Christoph Amma, and Tanja Schultz. 2015. Design and evaluation of a self-correcting gesture interface based on error potentials from EEG. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3375–3384. DOI: https://doi.org/10.1145/2702123.2702184

[108] Felix Putze, Jutta Hild, Rainer Kärgel, Christian Herff, Alexander Redmann, Jürgen Beyerer, and Tanja Schultz. 2013. Locating user attention using eye tracking and EEG for spatio-temporal event selection. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. ACM, New York, NY, 129–136. DOI: https://doi.org/10.1145/2449396.2449415

[109] Felix Putze, Johannes Popp, Jutta Hild, Jürgen Beyerer, and Tanja Schultz. 2016. Intervention-free selection using EEG and eye tracking. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, 153–160. DOI: https://doi.org/10.1145/2993148.2993199

[110] Felix Putze, Maximilian Scherer, and Tanja Schultz. 2016. Starring into the void? Classifying internal vs. external attention from EEG. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, New York, NY, 4 pages. DOI: https://doi.org/10.1145/2971485.2971555

[111] Felix Putze and Tanja Schultz. 2014. Investigating intrusiveness of workload adaptation. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, New York, NY, 275–281. DOI: https://doi.org/10.1145/2663204.2663279

[112] Felix Putze, Maik Schünemann, Tanja Schultz, and Wolfgang Stuerzlinger. 2017. Automatic classification of autocorrection errors in predictive text entry based on EEG and context information. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, 137–145. DOI: https://doi.org/10.1145/3136755.3136784

[113] Susanne Putze, Robert Porzel, Gian-Luca Savino, and Rainer Malaka. 2018. A manageable model for experimental research data: An empirical study in the materials sciences. In *Proceedings of the Advanced Information Systems Engineering*. John Krogstie and Hajo A. Reijers (Eds.), Lecture Notes in Computer Science, Springer International Publishing, 424–439.

[114] Melissa Quek, Daniel Boland, John Williamson, Roderick Murray-Smith, Michele Tavella, Serafeim Perdikis, Martijn Schreuder, and Michael Tangermann. 2011. Simulating the feel of brain-computer interfaces for design, development and social interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 25–28. DOI: https://doi.org/10.1145/1978942.1978947

[115] Richard Ramchurn, Sarah Martindale, Max L. Wilson, and Steve Benford. 2019. From director's cut to user's cut: To watch a brain-controlled film is to edit it. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 14 pages. DOI: https://doi.org/10.1145/3290605.3300378

[116] Elena Ratti, Shani Waninger, Chris Berka, Giulio Ruffini, and Ajay Verma. 2017. Comparison of medical and consumer wireless EEG systems for use in clinical trials. *Frontiers in Human Neuroscience* 11 (2017), 398. DOI: https://doi.org/10.3389/fnhum.2017.00398

[117] Héctor Rieiro, Carolina Diaz-Piedra, José Miguel Morales, Andrés Catena, Samuel Romero, Joaquin Roca-Gonzalez, Luis J. Fuentes, and Leandro L. Di Stasi. 2019. Validation of electroencephalographic recordings obtained with a consumer-grade, single dry electrode, low-cost device: A comparative study. *Sensors* 19, 12 (2019), 2808. DOI: https://doi.org/10.3390/s19122808

[118] Mathieu Rodrigue, Jungah Son, Barry Giesbrecht, Matthew Turk, and Tobias Höllerer. 2015. Spatio-temporal detection of divided attention in reading applications using EEG and eye tracking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 121–125. DOI: https://doi.org/10.1145/2678025.2701382

[119] Jean-Baptiste Sauvan, Anatole Lécuyer, Fabien Lotte, and Géry Casiez. 2009. A performance model of selection techniques for P300-based brain-computer interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 2205–2208. DOI: https://doi.org/10.1145/1518701.1519037

[120] A. Secerbegovic, S. Ibric, J. Nisic, N. Suljanovic, and A. Mujcic. 2017. Mental workload vs. stress differentiation using single-channel EEG. In *Proceedings of the CMBEBIH 2017*. Springer, 511–515.

[121] Pradeep Shenoy and Desney S. Tan. 2008. Human-aided computing: Utilizing implicit human processing to classify images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 845–854. DOI: https://doi.org/10.1145/1357054.1357188

[122] Yathunanthan Sivarajah, Eun-Jung Holden, Roberto Togneri, Greg Price, and Tele Tan. 2014. Quantifying target spotting performances with complex geoscientific imagery using ERP P300 responses. *International Journal of Human-Computer Studies* 72, 3 (2014), 275–283.

[123] Daniel Sjölie, Kenneth Bodin, Eva Elgh, Johan Eriksson, Lars-Erik Janlert, and Lars Nyberg. 2010. Effects of interactivity and 3d-motion on mental rotation brain activity in an immersive virtual environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 869–878. DOI: https://doi.org/10.1145/1753326.1753454

[124] Filip Škola and Fotis Liarokapis. 2019. Examining and enhancing the illusory touch perception in virtual reality using non-invasive brain stimulation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 247:1–247:12. DOI: https://doi.org/10.1145/3290605.3300477

[125] Mohammad Soleymani, Frank Villaro-Dixon, Thierry Pun, and Guillaume Chanel. 2017. Toolbox for emotional feature extraction from physiological signals (TEAP). *Frontiers in ICT* 4 (2017), 1. DOI: https://doi.org/10.3389/fict.2017.00001

[126] Erin Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. 2012. Brainput: Enhancing interactive systems with streaming fNIRS brain input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 2193–2202. DOI: https://doi.org/10.1145/2207676.2208372

[127] Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M. Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert J.K. Jacob. 2009. Using fNIRS brain sensing in realistic HCI settings: Experiments and guidelines. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 157–166. DOI: https://doi.org/10.1145/1622176.1622207

[128] Erin Treacy Solovey, Francine Lalooses, Krysta Chauncey, Douglas Weaver, Margarita Parasi, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, Paul Schermerhorn, Audrey Girouard, and Robert J.K. Jacob. 2011. Sensing cognitive multitasking for a brain-based adaptive user interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 383–392. DOI: https://doi.org/10.1145/1978942.1978997

[129] Erin T. Solovey and Felix Putze. 2021. Improving HCI with brain input: Review, trends, and outlook. *Foundations and Trends® in Human-Computer Interaction* 13, 4 (2021), 298–379. DOI: https://doi.org/10.1561/1100000078

[130] Daniel Szafir and Bilge Mutlu. 2012. Pay attention! Designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 11–20. https://doi.org/10.1145/2207676.2207679

[131] Daniel Szafir and Bilge Mutlu. 2013. ARTFul: Adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1001–1010. DOI: https://doi.org/10.1145/2470654.2466128

[132] Naoto Terasawa, Hiroki Tanaka, Sakriani Sakti, and Satoshi Nakamura. 2017. Tracking liking state in brain activity while watching multiple movies. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, 321–325. DOI: https://doi.org/10.1145/3136755.3136772

[133] Thomas Terkildsen and Guido Makransky. 2019. Measuring presence in video games: An investigation of the potential use of physiological measures as indicators of presence. *International Journal of Human-Computer Studies* 126 (2019), 64–80. DOI: https://doi.org/10.1016/j.ijhcs.2019.02.006

[134] Chi Thanh Vi, Kasper Hornbæk, and Sriram Subramanian. 2017. Neuroanatomical correlates of perceived usability. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, 519–532. DOI: https://doi.org/10.1145/3126594.3126657

[135] Keith F. Tipton, Richard N. Armstrong, Barbara M. Bakker, Amos Bairoch, Athel Cornish-Bowden, Peter J. Halling, Jan-Hendrik Hofmeyr, Thomas S. Leyh, Carsten Kettner, Frank M. Raushel, Johann Rohwer, Dietmar Schomburg, and Christoph Steinbeck. 2014. Standards for reporting enzyme data: The STRENDA consortium: What it aims to do and why it should be helpful. 1, 1 (2014), 131–137. DOI: https://doi.org/10.1016/j.pisc.2014.02.012

[136] Erin Treacy Solovey, Daniel Afergan, Evan M. Peck, Samuel W. Hincks, and Robert J. K. Jacob. 2015. Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fNIRS. *ACM Transactions on Computer-Human Interaction* 21, 6 (2015), 27 pages. DOI: https://doi.org/10.1145/2687926

[137] Anna Elisabeth van't Veer and Roger Giner-Sorolla. 2016. Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology* 67 (2016), 2–12.

[138] Boris M. Velichkovsky and John Paulin Hansen. 1996. New technological windows into mind: There is more in eyes and brains for human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 496–503. DOI: https://doi.org/10.1145/238386.238619

[139] Chi Vi and Sriram Subramanian. 2012. Detecting error-related negativity for interaction design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 493–502. DOI: https://doi.org/10.1145/2207676.2207744

[140] Chi Thanh Vi, Izdihar Jamil, David Coyle, and Sriram Subramanian. 2014. Error related negativity in observing interactive tasks. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3787–3796. DOI: https://doi.org/10.1145/2556288.2557015

[141] Chi Thanh Vi, Izdihar Jamil, David Coyle, and Sriram Subramanian. 2014. Error related negativity in observing interactive tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3787–3796. DOI: https://doi.org/10.1145/2556288.2557015

[142] Filip Škola and Fotis Liarokapis. 2019. Examining and enhancing the illusory touch perception in virtual reality using non-invasive brain stimulation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 12 pages. DOI: https://doi.org/10.1145/3290605.3300477

[143] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI research artifacts: Results of a self-reported survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[144] Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7 (2016), 1832.

[145] John Williamson, Roderick Murray-Smith, Benjamin Blankertz, Matthias Krauledat, and K-R Müller. 2009. Designing for uncertain, asymmetric control: Interaction design for brain–computer interfaces. *International Journal of Human-Computer Studies* 67, 10 (2009), 827–841.

[146] Katherine Wolstencroft, Stuart Owen, Olga Krebs, Quyen Nguyen, Natalie J. Stanford, Martin Golebiewski, Andreas Weidemann, Meik Bittkowski, Lihua An, David Shockley, Jacky L. Snoep, Wolfgang Mueller, and Carole Goble. 2015. SEEK: A systems biology data and model management platform. *BMC Systems Biology* 9, 1 (2015), 33. DOI: https://doi.org/10.1186/s12918-015-0174-y

[147] Shuo Yan, GangYi Ding, Hongsong Li, Ningxiao Sun, Yufeng Wu, Zheng Guan, Longfei Zhang, and Tianyu Huang. 2016. Enhancing audience engagement in performing arts through an adaptive virtual environment with a brain-computer interface. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, New York, NY, 306–316. DOI: https://doi.org/10.1145/2856767.2856768

[148] Mingqing Yang, Li Lin, and Slavko Milekic. 2018. Affective image classification based on user eye movement and EEG experience information. *Interacting with Computers* 30, 5 (2018), 417–432.

[149] Koji Yatani. 2016. Effect sizes and power analysis in hci. In *Proceedings of the Modern Statistical Methods for HCI*. Springer, 87–110.

[150] Yipeng Yu, Dan He, Weidong Hua, Shijian Li, Yu Qi, Yueming Wang, and Gang Pan. 2012. FlyingBuddy2: A brain-controlled assistant for the handicapped. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, New York, NY, 669–670. DOI: https://doi.org/10.1145/2370216.2370359

[151] Beste F. Yuksel, Michael Donnerer, James Tompkin, and Anthony Steed. 2010. A novel brain-computer interface using a multi-touch surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 855–858. DOI: https://doi.org/10.1145/1753326.1753452

[152] Beste F. Yuksel, Kurt B. Oleson, Lane Harrison, Evan M. Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn Piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5372–5384. DOI: https://doi.org/10.1145/2858036.2858388

[153] Thorsten O Zander, Jonas Brönstrup, Romy Lorenz, and Laurens R Krol. 2014. Towards BCI-based implicit control in human–computer interaction. In *Proceedings of the Advances in Physiological Computing*. Springer, 67–90.

[154] Thorsten O. Zander, Marius David Klippel, and Reinhold Scherer. 2011. Towards multimodal error responses: A passive BCI for the detection of auditory errors. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, New York, NY, 53–56. DOI: https://doi.org/10.1145/2070481.2070493

[155] Thorsten O Zander, Christian Kothe, Sabine Jatzev, and Matti Gaertner. 2010. Enhancing human-computer interaction with input from active and passive brain-computer interfaces. In *Proceedings of the Brain-computer interfaces*. Springer, 181–199.

[156] Rosanne Zerafa, Tracey Camilleri, Owen Falzon, and Kenneth P. Camilleri. 2018. A comparison of a broad range of EEG acquisition devices – is there any difference for SSVEP BCIs? *Brain-Computer Interfaces* 5, 4 (2018), 121–131. DOI: https://doi.org/10.1080/2326263X.2018.1550710

[157] Feng Zhou, Xingda Qu, Martin G Helander, and Jianxin Roger Jiao. 2011. Affect prediction from physiological measures via visual stimuli. *International Journal of Human-Computer Studies* 69, 12 (2011), 801–819.

[158] Feng Zhou, Xingda Qu, Jianxin Jiao, and Martin G Helander. 2014. Emotion prediction from physiological signals: A comparison study between visual and auditory elicitors. *Interacting with computers* 26, 3 (2014), 285–302.

Received August 2020; revised September 2021; accepted October 2021