Show me How You Use Your Mouse and I Tell You How You Feel? Sensing Affect with the Computer Mouse

Paul Freihaut and Anja S. Göritz

Abstract— Computer mouse tracking is a simple and cost-efficient way to gather continuous behavioral data. As theory suggests a relationship between affect and sensorimotor processes, the computer mouse might be usable for affect sensing. However, the processes underlying a connection between mouse usage and affect are complex, hitherto empirical evidence is ambiguous, and the research area lacks longitudinal studies. The present work brings forward a longitudinal field study in which 179 participants hourly self-reported their affect while their mouse usage was tracked both during their self-directed, contextless as well as task-bound computer use over the course of 14 days, resulting in a dataset comprising 10,760 instances of data collection. Extensive statistical analysis using null hypothesis significance testing and machine learning reveal weak and sporadic relationships between mouse usage and longitudinal self-reported affect at best. The results of this study challenge the immediate use of computer mouse tracking for longitudinal affect sensing and point to a necessity for more research.

Index Terms— affective computing, affect, measurement, computer mouse, field study, nonverbal signals

1 Introduction

THE computer mouse is a commonplace sensor in daily human-computer interaction. Tracking computer mouse usage conveniently and unobtrusively captures a rich stream of behavioral data without the need for sophisticated equipment and without requiring the user to change their customary behavior [1]. It comes as a surprise that only recently researchers have begun to explore the potential of computer mouse usage data in more detail, mostly to elucidate cognitive processes [2], [3]. Originating in the idea of affective computing, the present study seeks to explore the feasibility of mouse tracking as a tool for affect measurement. Leveraging the advantages of this sensing approach, computer mouse tracking might provide a useful addition to established, but often more cumbersome, methods of affect measurement [4], [5], and could contribute to practical applications as well as theoretical advances in the study of affect.

2 Linking the Use of the Computer Mouse and Affect

Zimmermann and colleagues [6] first suggested the potential of mouse tracking in affective computing. Their rationale was that affect causes distinctive and observable patterns in the way a person interacts with the computer mouse. Despite the intuitive appeal to this rationale, untangling the relationship between affect and mouse usage, however, is not straightforward, as the underlying processes are complex [7], [8], [9], [10].

A typical mouse usage action, such as navigating to a button and clicking on it, represents a goal-directed sensorimotor action. Current research indicates a connection between these sensorimotor actions and affect [11]. Theories generally attribute this relationship to either cognitive function or neuromotor pathways. The cognitive function pathway postulates that affect influences executive functions such as attentional control or working memory [12], [13], which are crucial in planning and controlling of motor actions [14], [15], [16]. The neuromotor pathway postulates that affect influences neuromotor processes such as corticospinal excitability [17], motor evoked potentials or muscle activity [18], which are crucial for motor actions. A growing body of studies support a relationship between affect and different movement attributes, such as the speed, accuracy, and variability of motor actions during task execution [19], [20], [21], [22], [23]. However, there is little theory that directly links affect and mouse usage. Most previous studies in the research area are propped on intuition that mouse usage signals affect as well as on practical reasons.

3 Empirical Studies on the Use of the Computer Mouse for Affect Sensing

Despite the fact that Zimmerman et al.'s proposal [6] to use the mouse for affect measurement has been around for almost 20 years, the empirical evidence remains sparse. Almost all studies are cross-sectional laboratory experiments that include affect manipulation and standardized mouse usage tasks. The results of these studies do not lend themselves to clear interpretation. Most studies reported findings in support of a relationship between affect and mouse usage, pointing out the potential of using mouse tracking for affect measurement (cf. [25], [26], [27], [28], [29]). However, there were also studies that did not identify a reliable

• *The authors are with the Chair of Behavioral Health Technology, University of Augsburg, Augsburg, Bavaria, Germany.*
E-mail: pfreihaut@gmail.com, anja.goeritz@uni-a.de

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

relationship (cf. [30], [31]). Adding to this inconsistency, some studies' small sample sizes and methodological limitations, such as confounding affect manipulation and mouse usage tasks, further muddled the interpretability of results [25]. Freihaut and colleagues [31] further critiqued the scarcity of open science practices. Transparently sharing data and data analysis code is crucial in this research area, because there is a lack of standardized methodological approaches (e.g., mouse data can be processed in many ways), which promotes finding and reporting unreliable outcomes.

Despite the capabilities of computer mouse tracking for continuous, long-term, and personalized data collection, longitudinal data remain by and large absent [1]. An exception being a field study that monitored mouse usage of 70 employees during their regular computer use at work over seven weeks [26]. The study findings suggest that in a state of stress (i.e., negative valence and high arousal), participants move their mouse at a higher speed at the cost of a decrease in accuracy or vice-versa, that is, a speed-accuracy trade-off, compared to a non-stress state.

4 The Present Study

The current state of research suggests a potential relationship between affect and mouse usage. However, we also highlighted the need for further investigation given the complex underlying processes, the lack of a solid theoretical foundation, ambiguities in the empirical evidence, as well as an almost complete lack of longitudinal studies. Utilizing the computer mouse as a fingerprint of the user's affect goes beyond finding sporadic significant relationships. To serve as a diagnostic marker of affect, mouse usage needs to correspond to affect in reliable and predictable ways [32].

This study contributes to filling this research gap by offering longitudinal data, as longitudinal research might be the most promising approach to study the feasibility of using computer mouse tracking in affective computing [32]. Our work is guided by the research question of whether there exists a systematic relationship between affect and mouse usage during everyday computer use. Moreover, we aim to address the proposed promise that mouse usage allows to reliably infer individuals' affect during their everyday computer use.

In the study, we used ecological momentary assessment (EMA) to collect mouse usage data and self-reported affect from participants multiple times a day over two weeks. EMA offers assessment in people's daily life [33, 34], which allows to evaluate the practical applicability of the measurement approach. In contrast to previous studies, we captured mouse usage during both, participants contextless regular computer use, as well as during a standardized task. We conceptualized affect within the core affect model [35], delineating it into two dimensions: valence (positive/negative feelings) and arousal (levels of excitement or calmness). Participants' self-reported ratings served as the ground truth for their affective states. Importantly, we regularly prompted participants to rate their current affective state, thereby capturing their moment-to-moment feelings of positivity/negativity and excitement/calmness. Yet, these ratings also partially reflect a participant's trait affect, representing their usual affect level across different situations and over time [36]. Our longitudinal approach enables us to disentangle state affect from trait affect and to consider both independently [37].

The data analysis followed a data-driven exploratory approach as well as open-science principles. Conducting a transparent and systematic empirical evaluation may best catalyze theoretical advancements and methodological standardization in this field.

5 Method

5.1 Design

The study was delivered via a "Study-App", which participants installed on their computer. Over 14 days, the app automatically initiated an instance of data collection once per hour. Each instance of data collection comprised three parts: (1) The Study-App discreetly tracked the position of the mouse cursor while the participant was engaged in their regular computer activities for a duration of five minutes (i.e., contextless mouse usage). (2) The Study-App prompted participants to complete a mouse usage task. (3) Participants were requested to report their current valence and arousal levels.

The study intentionally avoided an active affect manipulation as we aimed to capture natural variations in valence and arousal during everyday computer usage.

5.2 Participants

The study encompassed 179 participants who together completed a total of 10,760 instances of hourly data collection (per participant Mean = 60.11, SD = 40.09, Min = 0, Max = 224). This entails participants completing the mouse task, rating their valence and arousal, and having their self-directed mouse usage recorded. Participants were in part recruited via social media and word-of-mouth (convenience sample, n = 44), and in part via WisoPanel (panel sample, n = 135), an online access panel with participants from all walks of life [38], [39]. Table 1 shows sociodemographics. In the convenience sample, a higher percentage of participants are in the younger age groups. Given the proof-of-concept nature of this study, we decided to collaps the two samples. The number of participants in each age group is reasonably balanced in the combined dataset.

As part of an independent inquiry into the effect of remuneration on study participation, invited panel participants (N = 990) were randomly offered either 5€, 10€ or no remuneration for their participation. The convenience sample did not receive any remuneration. The two samples as well as the variation of remuneration were deemed to enhance the robustness of the present study's results.

Participation required the use of a physical computer mouse. Individuals who primarily used a trackpad, touch or another non-mouse computer input device were requested to abstain from participating. The app was available for Windows 10 (91.6%) and MacOS (8.4%).

TABLE 1
Sample Sociodemographics

	Total Sample		*Panel Sample*		*Conv. Sample*	
	N	*%*	*N*	*%*	*N*	*%*
Age						
< 30	41	22.9	10	7.4	31	70.5
30 – 39	26	14.5	18	13.3	8	18.2
40 – 49	31	17.3	29	21.5	2	4.5
50 – 59	33	18.4	31	23.0	2	4.5
>= 60	38	21.3	38	28.1	0	0.0
Not reported	10	5.6	9	6.7	1	2.3
Gender						
Male	94	52.5	72	53.3	22	50.0
Female	83	46.4	61	47.4	22	50.0
Not reported	2	1.1	2	1.5	0	0.0
Hand to use the mouse						
Right	170	95.0	127	94.1	43	97.7
Left	8	4.5	7	5.2	1	2.3
Not reported	1	0.5	1	0.7	0	0.0
Study remuneration						
0€	74	41.34	30	22.2	44	100.0
5€	45	25.14	45	33.3	0	0.0
10€	60	33.52	60	44.4	0	0.0

5.3 Measures

5.3.1 Contextless Mouse Usage

In the initial part of each instance of data collection, the Study-App recorded a participant's self-directed mouse usage behavior during their regular computer use during a 5 min interval. The purpose of this was to capture a snapshot of the participant's natural and unconstrained mouse behavior.

The recording of contextless mouse data was time-based. The mouse cursor's x- and y-position were logged on the entire computer screen along with a timestamp at a sampling rate of 50 Hz (i.e., one datapoint every 20 ms). As a result, a five-minute segment of recorded mouse use yielded up to 15,000 raw cursor position data points. The sampling frequency was chosen as a compromise between sampling accuracy and size of the recorded dataset. Moreover, the sampling rate is similar to the event-based sampling approach, which was used to capture task-specific mouse usage. The recording ended once the participant commenced the subsequent mouse-usage task. Fig. 2 shows an example of the contextless mouse data.

Fig. 1. Example of recorded mouse movement during the 5-min contextless computer use interval. The dots represent the recorded x- and y-positions on the screen. The colors mark individual movement episodes. Movement episodes are delimited by pauses exceeding a specified threshold (e.g., 1 sec).

5.3.2 Task-Specific Mouse Usage

The second phase of each data collection instance was a simple point-and-click task (Fig. 1). Participants were presented with a 4-by-4 grid of solid circles and instructed to click on 7 out of the 16 circles in a specified sequence. The first circle to be clicked was highlighted. Upon being clicked, the circle was marked as 'clicked' and the next circle in the sequence was highlighted. This continued until all 7 circles had been clicked. The goal of the point-and-click task was to track participants mouse usage in a standardized way during a prototypical mouse usage task [29]. To limit task habituation, the click sequence for each instance of the task was randomly selected from a prepared set of 25 sequences.

During the task, the app logged all mouse usage behavior inside the app's task window. Data were collected in an event-based manner, meaning a data point was generated each time a mouse event (i.e., positional change or click) occurred. The sampling rate of continuous mouse movement was around 50 Hz. Each data point consisted of the name of the mouse event, the cursor's x/y position in the task window, a timestamp and the number of circles clicked so far. The median count of collected raw mouse usage datapoints in the point-and-click task was 235.

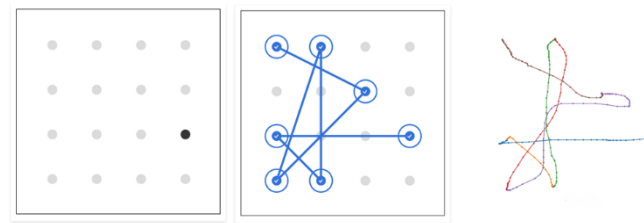

Fig. 2. Point-and-click task. The left panel shows the beginning and the middle panel shows the ending of the point-and-click task. The right-hand panel shows an example of the logged mouse usage data during the task. The dots alongside the mouse movement path represent the recorded x- and y-positions on the screen. The colors mark movement episodes (i.e., the mouse movement between consecutive circles shown in the middle panel).

5.3.3 Affect Measurement

Participants reported their affect through two questions pertaining to their current feelings of valence (ranging from negative to positive) and arousal (ranging from excited to calm). The responses were captured using a slider scale that ranged from 0 to 100, with the default position set at 50. Valence and arousal are common measures of affect, for example, via the SAM [40]. By limiting affect assessment to two questions per data collection, we tried to minimize the burden placed on participants. Single item measurements have shown acceptable predictive validity as compared to multiple-item measures, which make them especially attractive in intensive longitudinal designs [41]. We did not ask participants directly about potential external factors that might influence affect, such as caffeine usage. However, we logged the time of the day and date of each measurement.

5.4 Procedure

The Study-App automatically handled the study procedure without the need for intervention either by

participants or the experimenter, thus ensuring an objective data collection process. Before the start of the data collection, participants had to complete an introductory tutorial. The study ended 14 days after the start of the data collection (see Supplement 1 for details). After each data collection instance, all collected data were anonymously saved in a database. No data were saved when participants aborted or opted to skip a data collection instance.

The app window was square-shaped, with its height and width set to 80% of the available screen height. It appeared in the center of the primary computer screen. Participants could not resize the app window, but they were able to change the position of the window by dragging it. The app was programmed with Electron.js and used React.js for building the user interface.

5.5 Data Analysis

Recall that the study dataset contained 10,760 instances of data collections from 179 participants. Each data collection instance included contextless mouse data, contextual mouse-task data, and self-reported affect ratings. This allowed us to independently explore the relationship between affect and mouse usage in both contexts. The data analysis procedure entailed two steps: preprocessing the mouse usage data followed by statistical analysis. An illustration of this process is provided in Figure 3.

5.5.1 Data Preprocessing

5.5.1.2 **Contextless Mouse Data:** It involved three preprocessing stages (details in Supplement 2):

1) Data quality inspection: The raw data were vetted for quality, with each data collection instance expected to contain up to 15,000 mouse cursor position data points. Seventeen instances with recording errors were removed, as were the data of 9 participants with less than 3 valid data collections. The refined dataset included 10,735 data collections from 170 participants (Mean = 63.14, Median = 59; SD = 38.59, Min = 7, Max = 224).
2) Feature creation: The raw data were transformed into discrete mouse usage features. In line with the procedure in [26], the dataset was divided into periods of mouse movement and non-movement. A movement period commenced with a change in mouse position and ended when no positional changes were detected for a specified threshold. To account for the lack of a generally agree-upon threshold, we created three datasets based on thresholds of 1 sec, 2 sec, and 3 sec. For each pause threshold dataset, we calculated 31 spatial and temporal mouse usage features in accordance with the mouse tracking literature and available mouse data processing software [42], [3]. See Table 2 for an overview of the features, for details see the supplement.
3) Feature reduction: Highly correlated mouse usage features ($r > .8$) were removed from each dataset to decrease redundancy.

Following preprocessing, we obtained three distinct contextless mouse datasets: $D_{1\text{-sec-pause-thresh}}$, $D_{2\text{-sec-pause-thresh}}$, and $D_{3\text{-sec-pause-thresh}}$. Independently analyzing each dataset represents a multiverse analysis, which benefits robustness and transparency [43]. Notwithstanding, given the virtually limitless alternatives in data preprocessing, our choices were ultimately a compromise between exploring various reasonable preprocessing scenarios and managing computational demands and result complexity.

5.5.1.1 **Mouse-Task Data:** It involved five preprocessing stages (details in Supplement 3):

1) Data quality inspection: The raw data were vetted for quality, with each data collection instance containing a median of 235 raw data points. Twenty-two instances with recording errors were removed, as were the data of 10 participants with less than 3 valid data collections. The refined dataset included 10,729 data collections from 169 participants (Mean = 63.49, Median = 60; SD = 38.60, Min = 7, Max = 224).
2) Feature creation: The raw data were transformed into 41 spatial and temporal mouse usage features. See Table 2 for an overview of the features, for details see the supplement.
3) Outlier removal: The mouse usage features were checked for anomalies (e.g., random mouse movements instead of straight paths between click points). Given the absence of a agreed-upon procedure to identify careless responders, and careless responding might carry information about affect, we created three datasets with different outlier removal procedures. In the first dataset, we removed two cases with a task duration exceeding 15 min

TABLE 2
Summary of the Mouse-Usage Features

Feature Category	Description
Contexless Mouse-Usage Features	
Speed (12 features)	Describe average and variation in mouse speed, acceleration and jerk during mouse movement episodes
Distance & Accuracy (11 features)	Describe average and variation in mouse distance, directional changes and angles between consecutive mouse movement vectors during the mouse movement episodes
Duration (4 features)	Describe average and variation in the movement episode duration and the time of no movement
Other (4 features)	Describe the number of movement episodes, total recording time, number of lockscreen episodes and lockscreen time
Mouse-Task Features	
Speed (18 features)	Describe average and variation in speed, acceleration and jerk of during the mouse task as well as during the mouse task trials
Distance & Accuracy (17 features)	Describe average and variation in mouse distance, directional changes, angles between consecutive mouse movement vectors, and the distance from an ideal task movement during the mouse task as well as during the mouse task trials
Duration & Reaction time (5 features)	Describe average and variation in the duration of the mouse task trials and the reaction time in each mouse task trial, which is the time difference between the start of a trial and the first movement towards the target
Clicks (1 feature)	Describe the number of mouse clicks during the mouse task

(median task duration = 6.97 sec). In the second and third dataset, we removed outliers using the interquartile range (IQR) method with thresholds of 2.5 and 3.5.

4) Click order harmonization: Potential systematic differences in mouse usage features between task click orders were harmonized using linear equating [44].
5) Feature reduction: Highly correlated features (r > .8) were removed from each dataset.

Following preprocessing, we obtained three distinct mouse-task datasets: $D_{dur.\text{-}cutoff}$, $D_{IQR\text{-}2.5}$, and $D_{IQR\text{-}3.5}$. Again, the data preprocessing steps highlight the researcher degrees of freedom when working with mouse usage data, and the selected datasets do not cover all possible preprocessing options.

5.5.2 *Statistical Analysis*

Our analysis aimed to explore the bivariate relationship between mouse usage and affect (valence and arousal) and assess the feasibility of reliably inferring affect from mouse movements during everyday computer use. Two analytical approaches have commonly been employed in similar research:

1) Null hypothesis significance testing (NHST): NHST enables population inferences about the links between mouse usage features and affect. For instance, [26] used Bayesian mixed-model logistic regression to test if stress is characterized by a speed-accuracy trade-off in mouse movements.
2) Machine Learning (ML): ML tests if mouse usage successfully predicts affect. For instance, [25] utilized random forest regression to predict varying feeling states from 16 mouse usage features.

Both data analysis approaches offer unique advantages: NHST helps to uncover the underlying processes of the relationship between affect and mouse usage behavior, while ML helps to evaluate the reliability of affect prediction from mouse usage [45]. To support multifinality, our exploratory data analysis included both, NHST as well as ML.

Note that our exploratory research approach allows building and testing an infinite number of statistical models (e.g., testing interaction effects between mouse usage features). Given the emerging state of this research area, we focused on relatively simple models of the relationship between mouse usage and affect. This approach best addresses the fundamental proposal that mouse usage and affect are reliably related, while also managing the complexity of the statistical analysis. Both the NHST and ML were implemented similarly for testing the relationship between contextless mouse usage and affect as well as between contextual mouse-task mouse usage and affect.

5.5.2.1 **NHST Analysis:** We used linear mixed models to test the relationship between single mouse usage features (independent variable) and affect (valence or arousal – dependent variable). For each affect measure and mouse usage feature, we compared three models.

Null model: The model included a random intercept for each participant, but no predictor variable. The null model`s intraclass correlation (ICC) informs on how much variance in the outcome variable is due to variation between persons (i.e., individual differences in the average affect across the measurements; trait affect) and variation within persons (i.e., measurement-specific deviations in affect from one's usual level; state affect). The null model also serves as a baseline to assess how much additional information each mouse usage features provides. Formally, the null model is defined as

$$A_{ij} = \beta_0 + \beta_{0i} + \varepsilon_{ij} \quad (1)$$

where participant i's affect (valence or arousal) at measurement j, A_{ij}, is a function of the overall intercept, β_0, participant-specific variation in the intercept, β_{0i}, and error, ε_{ij}.

Fixed effect model: We added one mouse usage feature into the model as a fixed effect predictor. Like affect, mouse usage data contain both, between-person variation (i.e., trait mouse usage) and within-person variation (i.e., state mouse usage). Each source of variation can exert its own

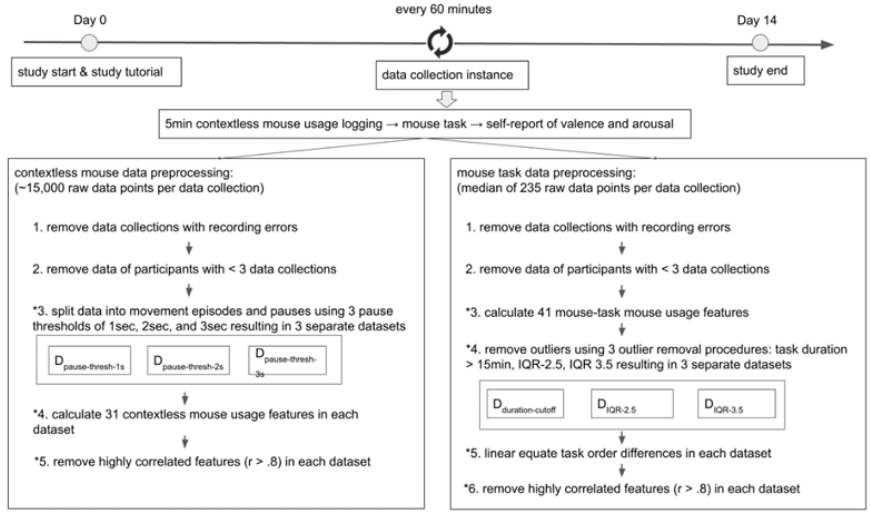

Fig 3. Overview of the study procedure and data preprocessing. In the machine learning analysis; *steps were conducted using the training dataset only on the test dataset.

effect on the outcome [37]. The between-person effect represents a trait effect, and the within-person effect represents a state effect. We included both sources of variance into the model by splitting the mouse usage predictor into a between-person (trait) and a within-person (state) predictor using person-mean-centering [37]:

$$Trait\ Predictor: MouseFeature_{Trait_i} = \overline{MouseFeature_i} \quad (2)$$

$$State\ Predictor: MouseFeature_{State_{ij}} = MouseFeature_{ij} - \overline{MouseFeature_i} \quad (3)$$

where $MouseFeature_{ij}$ is the participant i's mouse usage feature at measurement j, and $\overline{MouseFeature_i}$ is participant i's mean mouse usage feature. Formally, the fixed effect model is defined as

$$A_{ij} = \beta_0 + \beta_1 MouseFeature_{Trait_i} + \beta_2 MouseFeature\_State_{ij} + \beta_{0i} + \varepsilon_{ij} \quad (4)$$

where participant i's affect (valence or arousal) at measurement j, A_{ij}, is a function of the overall intercept, β_0, the fixed slope parameter of the trait mouse usage feature, β_1, the fixed slope parameter of the state mouse usage feature, β_2, participant-specific variation in the intercept, β_{0i}, and error, ε_{ij}.

Random slope model: We allowed the state effect of mouse usage on affect to vary between participants (i.e., a random slope). The random-slope model considers potential participant-specific relationships between affect and the mouse usage feature (e.g., an increase in arousal might be associated with an increase in mouse speed for some participants, but with a decrease in mouse speed for other participants). The random slope model is defined as

$$A_{ij} = \beta_0 + \beta_1 MouseFeature_{Trait_i} + \beta_2 MouseFeature_{State_{ij}} + \beta_{3i} MouseFeature_{State_{ij}} + \beta_{0i} + \varepsilon_{ij} \quad (5)$$

where participant i's affect (valence or arousal) at measurement j, A_{ij}, is a function of the overall intercept, β_0, the fixed slope parameter of the trait mouse usage feature, β_1, the fixed slope parameter of the state mouse usage feature, β_2, the participant-specific slope parameter of the state mouse usage feature, β_{3i}, participant-specific variation in the intercept, β_{0i}, and error, ε_{ij}.

Considering that linear mixed models assume normally distributed residuals [46] and that neither the outcome variables nor most mouse usage features follow a normal distribution, we used rank-based inverse normalization [47] to transform these variables before their inclusion in the models. We estimated all models using maximum likelihood.

5.5.2.2 **ML Analysis:** We used random forest regression to test if affect (valence or arousal as either outcome variable) can be predicted from mouse usage (all mouse usage features of a dataset as input features). For each affect measure, we compared two ML models:

Null model: The model included a single input feature: the Participant ID number to predict affect. The inclusion of the Participant ID is akin to assigning each participant a unique intercept, hence the model predicts each individual's average affect level (i.e., the model accounts for the trait variance of affect). The null model serves as a baseline to later assess how much additional predictive information the mouse usage features provide.

Full model: The model included all mouse usage features as input features, along with the Participant ID. For both models, we trained them using the chronologically first 80% of each participant`s data (training dataset) and tested their performance on the remaining 20% data of each participant`s data (test dataset). This mimics the potential use case of personalized affect prediction. We chose the random forest algorithm because it has proven effective with mouse usage data in previous studies (c.f. [48], [28], [25]). The hyperparameters of the random forest were tuned with randomized grid search in a 5-fold cross validation loop [49]. To gain insight into each input feature's influence on the prediction performance, we computed permutational feature importance scores [50].

6 Results

We explored the relationship between affect and mouse usage in two contexts: during user-directed, contextless computer use and during a standardized, contextual mouse task. We transformed either raw data set—contextless mouse data and mouse-task data—into three distinct datasets and conducted analysis using both NHST and ML. Due to limited journal space and for the sake of clarity, this section focuses only on the crucial results. For a comprehensive review of all results, please refer to Supplements 4 (contextless mouse usage results) and 5 (mouse-task results).

6.1 Descriptive Statistics of Valence and Arousal

Across all data collection instances, participants reported their valence as more positive than negative (mean = 69.87, std = 23.26, range = 0 - 100 and their arousal as more calm than excited (mean = 69.28, std = 23.93, range = 0 - 100). The average within-participant standard deviations were 12.76 for valence and 15.12 for arousal. The distribution of valence and arousal responses is in Figure 4.

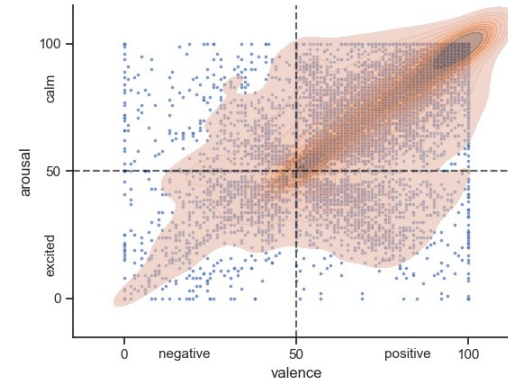

Fig. 4. Valence and arousal ratings across all instances of data collection. The scatterplot shows the individual ratings. The kernel density estimation (KDE) plot provides a smoothed representation of the data density, highlighting areas of higher and lower concentration of ratings.

TABLE 3
Contextless Mouse Usage NHST Analysis Result Summary

Outcome	Dataset	Dataset Characteristics		Null Model			FE Model						RS Model					
		N	# Features	R^2-cond	R^2-marg	R^2-cum	R^2-cond		R^2-marg		R^2-cum		R^2-cond		R^2-marg		R^2-cum	
							Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Valence	1s pause thresh	10735	16	0.645	0.000	0.640	0.645	0.647	0.020	0.069	0.640	0.640	0.647	0.650	0.020	0.071	0.644	0.647
	2s pause thresh	10735	14	0.645	0.000	0.640	0.645	0.647	0.018	0.055	0.640	0.640	0.647	0.651	0.018	0.056	0.644	0.647
	3s pause thresh	10735	15	0.645	0.000	0.640	0.645	0.647	0.017	0.058	0.640	0.640	0.647	0.651	0.017	0.057	0.644	0.648
Arousal	1s pause thresh	10735	16	0.561	0.000	0.565	0.561	0.564	0.016	0.056	0.565	0.566	0.564	0.569	0.017	0.060	0.569	0.573
	2s pause thresh	10735	14	0.561	0.000	0.565	0.561	0.563	0.014	0.042	0.565	0.566	0.564	0.569	0.016	0.045	0.569	0.573
	3s pause thresh	10735	15	0.561	0.000	0.565	0.561	0.563	0.014	0.044	0.565	0.566	0.564	0.568	0.014	0.045	0.570	0.574

Note. The table shows the average and maximum model explanatory power of the linear mixed effect models that were calculated for each mouse usage feature in each of the three datasets. The null model is a random intercept only model. The FE-models included the mouse usage feature as a fixed effect. The RS-models included the mouse usage feature as a fixed effect and as a random slope. R^2-cond (conditional R^2) quantifies the explanatory power of both, the fixed effects and the random effects. R^2-marg (marginal R^2) quantifies the explanatory power of the fixed effects. R^2-cum (cumulative R^2) quantifies the explanatory power as the square of the correlation between the model`s predicted outcome and the observed outcome.

6.2 Contextless Mouse Usage Results

Recall that there are three distinct datasets for contextless mouse usage, each representing a different movement pause threshold to distinguish between mouse movement and non-movement episodes: $D_{1\text{-sec-pause-thresh}}$, $D_{2\text{-sec-pause-thresh}}$, and $D_{3\text{-sec-pause-thresh}}$. All analysis were independently run for each dataset.

6.2.1 NHST Results

We compared three linear mixed models: (1) A null model with a random intercept, (2) a fixed effect model, which includes both the state and trait components of an individual mouse usage feature as fixed effects, and (3) a random slope model, which also includes the random slope for the state mouse usage feature. Our primary criterion for model evaluation was explanatory power. Specifically, we computed marginal R^2 (R^2-marg), conditional R^2 (R^2-cond), and cumulative R^2 (R^2-cum). R^2-marg quantifies the explanatory power of the fixed effects. R^2-cond quantifies the explanatory power of both, the fixed effects and the random effects, that is, the total model [51]. R^2-cum quantifies the explanatory power as the square of the correlation between the model`s predicted outcome and the observed outcome [37]. There were 16 mouse usage features in $D_{1\text{-sec-pause-thresh}}$, 14 features in $D_{2\text{-sec-pause-thresh}}$, and 15 features in $D_{3\text{-sec-pause-thresh}}$. We applied the Benjamini-Hochburg procedure to control the false discovery rate due to multiple testing [52].

Null model results: We computed a null model for each dataset and outcome variable. The average model Intraclass Correlation (ICC) indicates that 56% of the variance in arousal and 64% of the variance in valence was due to between-person mean differences (trait affect), while 44% and 37% of the variance was attributed to within-person variations (state affect).

Fixed effect model results: We computed a fixed effect model for each mouse usage feature in each dataset. This totaled 45 fixed effect models per outcome variable. Compared to the null model, 11 models (24%) for arousal and 18 models (40%) for valence had a significantly better fit. The average increase in explanatory power of all models compared to the null model was small: For arousal, there was an average increase of 0.00010 for R^2-cond (max = 0.0026), 0.015 for R^2-marg (max = 0.056), and 0.0000061 for R^2-cum (max = 0.00047). For valence, there was an average increase of 0.00013 for R^2-cond (max = 0.0022), 0.018 for R^2-marg (max = 0.069), and 0.000029 for R^2-cum (max = 0.00026).

Random slope model results: We computed a random slope model for each mouse usage feature in each dataset. This totaled 45 random slope models per outcome variable. Compared to the fixed effect model, 26 models (58%) for arousal and 31 models (69%) for valence had a significantly better fit. On average, the random slope models showed a slight increase in explanatory power as compared to the fixed effect models. However, the absolute explanatory power of all random slope models was small. For arousal, there was an average increase of 0.0027 for R^2-cond (max = 0.0061), 0.00042 for R^2-marg (max = 0.0043), and 0.0038 for R^2-cum (max = 0.0081). For valence, the average increase was 0.0024 for R^2-cond (max = 0.0055), 0.00042 for R^2-marg (max = 0.0025), and 0.0038 for R^2-cum (max = 0.0075).

6.2.2 Machine Learning Results

We compared two models: (1) a null model using the Participant ID as sole predictor, analogous to the random intercept model in the mixed model analysis, and (2) a full model, incorporating all mouse features alongside the Participant ID. To evaluate the models' predictive performance, we computed the coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE) between the predicted and observed outcome values in the test dataset. Training and testing data sizes remained consistent across all three datasets at N_{train} = 8,524 and N_{test} = 2,211, respectively. $D_{1\text{-sec-pause-thresh}}$ had 15 mouse usage features, $D_{2\text{-sec-pause-thresh}}$ had 12 features, and $D_{3\text{-sec-pause-thresh}}$ had 13 features. Note that the number of features varies between the NHST and ML analyses datasets, because in the ML analysis, to prevent data leakage into the test data the feature reduction was done using the training data only.

For arousal, null models displayed an average R^2 of .52, an MSE of 276.73, and an MAE of 11.49. The full model saw a decrease in average R^2 to .44 and increases in MSE to 328.56 and MAE to 13.47. The Participant ID emerged as the most important feature.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3357733

TABLE 4
Contextless Mouse Usage Machine Learning Analysis Result Summary

Outcome	Dataset	Dataset Characteristics			Null Model			Full Model		
		N-train	N-test	# Features	R^2	MSE	MAE	R^2	MSE	MAE
Valence	1s pause thresh	8524	2211	16	0.58	218.82	9.57	0.46	281.61	12.23
	2s pause thresh	8524	2211	13	0.58	219.48	9.58	0.47	273.54	11.94
	3s pause thresh	8524	2211	14	0.58	218.91	9.57	0.47	271.93	11.82
Arousal	1s pause thresh	8524	2211	16	0.52	276.92	11.49	0.43	333.80	13.72
	2s pause thresh	8524	2211	13	0.52	276.51	11.48	0.44	327.16	13.35
	3s pause thresh	8524	2211	14	0.52	276.77	11.48	0.44	324.71	13.35

Note. MSE = Mean Squared Error, MAE = Mean Absolute Error.

For valence, the null model displayed an average R^2 of .58, an MSE of 219.07, and MAE of 9.5. The full model saw a decrease in R^2 to .47, an increase in MSE to 275.69 and MAE to 12.00. Again, the Participant ID was the most important feature.

6.3 Mouse-Task Results

Recall that there are three distinct mouse-task datasets, each representing a different outlier removal procedure: $D_{dur.-cutoff}$, $D_{IQR-2.5}$, and $D_{IQR-3.5}$. All analyses were independently run for each dataset.

6.3.1 NHST Results

The NHST analysis of the mouse-task data mirrored the procedure employed with the contextless mouse data. We compared (1) a null model, (2) a fixed effect model and (3) a random slope model. The primary evaluation criterion was the model's exploratory power. $D_{dur-curoff}$ had N = 10,272 observations and 21 mouse usage features, $D_{IQR-2.5}$ had N = 8,761 observations and 22 mouse usage features, and $D_{IQR-3.5}$ had N = 9,526 observations and 20 mouse usage features. The false discovery rate was controlled with all significance tests.

Null model results: We computed a null model for each dataset and outcome variable. The ICC revealed that 56% of the arousal variance and 64% of the valence variance were due to between-person mean differences (trait affect). The within-person variation accounted for 44% of the arousal variance and 37% of the valence variance (state affect).

Fixed effect model results: We computed a fixed effect model for each mouse usage feature in every dataset, totaling 63 fixed effect models per outcome variable. In relation to the null model, 35 models (56%) for arousal showed a significantly superior fit. However, none of the models for valence demonstrated a significantly better fit compared to the null model. The average increase in explanatory power of all models compared to the null model was minor or even negative. For arousal, there was an average increase of 0.00014 for R^2-cond (max = 0.0011), 0.010 for R^2-marg (max = 0.034), and 0.00040 for R^2-cum (max = 0.0015). For valence, there was an average decrease of 0.0015 for R^2-cond (max = 0.00026), an average increase of 0.0067 for R^2-marg (max = 0.036), and an average increase of 0.00006 for R^2-cum (max = 0.00021).

Random slope model results: We computed a random slope model for each mouse feature in every dataset, which resulted in 63 random slope models per outcome variable. Compared to the fixed effect model, 30 models (48%) for arousal and 25 models (40%) for valence demonstrated a significantly superior fit. On average, the random slope models showed a slight increase in explanatory power as compared to the fixed effect models. However, the absolute explanatory power of all random slope models was small: For arousal, there was an average increase of 0.0026 for R^2-cond (max = 0.0064), 0.00030 for R^2-marg (max = 0.0025), and 0.0034 for R^2-cum (max = 0.010). For valence, the average increase was 0.0016 for R^2-cond (max = 0.0051), 0.000097 for R^2-marg (max = 0.0031), and 0.0075 for R^2-cum (max = 0.0075).

6.3.2 Machine Learning Results

The ML analysis of the mouse-task data mirrored the contextless mouse usage ML analysis. We compared (1) a

TABLE 5
Mouse Task NHST Analysis Result Summary

Outcome	Dataset	Dataset Characteristics		Null Model			FE Model						RS Model					
		N	# Features	R^2-cond	R^2-marg	R^2-cum	R^2-cond		R^2-marg		R^2-cum		R^2-cond		R^2-marg		R^2-cum	
							Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Valence	Dur. Cutoff	10272	21	0.645	0.000	0.640	0.644	0.644	0.007	0.034	0.640	0.640	0.645	0.649	0.007	0.037	0.643	0.647
	IQR 2.5	8761	22	0.645	0.000	0.640	0.641	0.642	0.006	0.033	0.634	0.635	0.643	0.645	0.005	0.020	0.637	0.639
	IQR 3.5	9526	20	0.645	0.000	0.640	0.645	0.646	0.007	0.035	0.637	0.637	0.647	0.648	0.006	0.018	0.639	0.642
Arousal	Dur. Cutoff	10272	21	0.559	0.000	0.565	0.559	0.560	0.010	0.023	0.565	0.566	0.562	0.566	0.010	0.023	0.569	0.576
	IQR 2.5	8761	22	0.554	0.000	0.555	0.554	0.555	0.010	0.034	0.556	0.557	0.557	0.561	0.008	0.018	0.559	0.565
	IQR 3.5	9526	20	0.558	0.000	0.559	0.558	0.559	0.010	0.027	0.560	0.560	0.560	0.565	0.010	0.027	0.563	0.569

Note. The table shows the average and maximum model explanatory power of the linear mixed effect models that were calculated for each mouse usage feature in each of the three datasets. The null model is a random intercept only model. The FE-models included the mouse usage feature as a fixed effect. The RS-models included the mouse usage feature as a fixed effect and as a random slope. R^2-cond (conditional R^2) quantifies the explanatory power of both, the fixed effects and the random effects. R^2-marg (marginal R^2) quantifies the explanatory power of the fixed effects. R^2-cum (cumulative R^2) quantifies the explanatory power as the square of the correlation between the model`s predicted outcome and the observed outcome.

TABLE 6
Mouse Task Machine Learning Result Summary

Outcome	Dataset	Dataset Characteristics			Null Model			Full Model		
		N-train	N-test	# Features	R^2	MSE	MAE	R^2	MSE	MAE
Valence	Dur. Cutoff	8520	2207	22	0.58	219.30	9.57	0.51	254.35	11.40
	IQR 2.5	6960	1784	23	0.57	222.37	9.60	0.50	258.39	11.23
	IQR 3.5	7573	1955	22	0.57	224.14	9.61	0.50	258.58	11.20
Arousal	Dur. Cutoff	8520	2207	22	0.52	277.17	11.49	0.46	317.02	13.33
	IQR 2.5	6960	1784	23	0.52	280.11	11.54	0.43	328.00	13.51
	IQR 3.5	7573	1955	22	0.51	281.14	11.54	0.44	324.81	13.50

Note. MSE = Mean Squared Error, MAE = Mean Absolute Error.

null model using the Participant ID as sole predictor, and (2) a full model, incorporating all mouse usage features alongside the Participant ID. Model performance was evaluated using R^2, MSE, and MAE. $D_{dur.\text{-}cutoff}$ was split into N_{train} = 8,520 observations and N_{test} = 2,207 observations and had 22 mouse features. $D_{IQR\text{-}2.5}$ was split into N_{train} = 6,960 observations and N_{test} = 1,784 observations and had 23 features. $D_{IQR\text{-}3.5}$ was split into N_{train} = 7,573 observations and N_{test} = 1,955 observations and had 22 features.

For arousal, the null models had an average R^2 of .52, an MSE of 279.47, and an MAE of 11.52. The full model saw a decrease in R^2 to .44 and increases in MSE to 323.28 and MAE to 13.44. The participant ID emerged as the most important feature.

For valence, the null model had an average R^2 of .60, an MSE of 221.94, and an MAE of 9.60. The full model saw a decrease in R^2 to .50 and increases in MSE to 257.11 and MAE to 11.28. Again, the participant ID was the most important feature.

7 Discussion

The computer mouse is a promising affect sensing approach, because it conveniently produces a rich stream of continuous behavioral data. This study explored the relationship between mouse usage and affect in a longitudinal field setting. Spanning 14 days, we hourly tracked mouse usage during participants' self-directed, contextless computer use as well as during a standardized, contextual mouse task. As ground truth of affect, participants rated their current feeling state's valence and arousal.

The interpretation of the results is complex, which is not unexpected considering the numerous statistical tests conducted. In the NHST analysis, there were tentative findings in support of a relationship between mouse usage and affect with both, the contextless mouse data and the mouse-task data. Several fixed effect models offered a significantly better fit than the null model, also after applying false discovery rate control. Furthermore, random slope models often fit better than the fixed effect models, implying potential person-specific relationships between mouse usage and affect. However, the incremental explanatory power across all non-null models over the null model was minimal, indicating a practically negligible correlation between mouse usage and affect. The machine learning results underscore this observation. The inclusion of mouse usage features in the ML models did not enhance affect prediction, but decreased prediction performance as compared to the null model. This decrease is likely due to overfitting on random noise arising from the inclusion of additional, non-informative input features.

Given this pattern of findings we refrain from highlighting specific mouse usage features as potential indicators of affect. However, we encourage further investigation of tentatively promising individual features outlined in the supplemental files. The pattern suggests that if mouse usage can indeed be a reliable indicator of affect, arousal might be more predictable than valence. Moreover, tracking mouse usage in standardized tasks might be more effective in predicting affect than tracking contextless mouse usage. In the mouse task, affect appeared more strongly correlated with state than with trait mouse features. By contrast, in contextless mouse usage, trait mouse features held a stronger correlation with affect than state mouse features. Nevertheless, the relationship between mouse usage and affect was, at best, marginal, regardless of whether mouse usage was contextless or contextualized.

Addressing the research questions of this study, our findings indicate a limited and uncertain link between mouse usage and affect. At this stage, it is premature to assert that everyday computer mouse usage can reliably predict individuals' affect. As such, the potential of the computer mouse as a tool in affect sensing should be regarded with skepticism [30], [31]. Such skepticism might be particularly pertinent in naturalistic settings. Promising results from laboratory settings do not necessarily translate into real-life settings [32]. For example, a recent study shows that heart rate variability (HRV), despite its effectiveness in lab settings, showed little predictive value for self-reported stress in an everyday life scenario [53]. This corresponds to our results. Such findings are important because they highlight the need for longitudinal studies outside of the laboratory.

Note that Banholzer and colleagues [26] drew a more positive conclusion from their longitudinal study data, which showed a significant relationship between everyday mouse usage and self-reported stress. However, although the authors suggested that their results indicate that the mouse could be used to predict the stress level of computer users, they did not specifically test such a prediction in their analysis. We bridged this gap by reanalyzing their data [54], which failed to provide reliable stress prediction on new data (see Supplement 6). Thus, the results of both studies might be more similar than their diverging

interpretations suggest.

From a theoretical perspective, the weak link between mouse usage and affect appears inconsistent with studies that link affect to motor control [11]. As previously discussed, one explanation could be the context-dependent nature of the relationship between mouse usage and affect in real-life settings [32], [55]. Most evidence that links affect to motor control is from laboratory studies with specific affect manipulation and isolated contexts. In contrast, we captured natural variations in affect in an everyday context and did not include contextual variables in our data analysis. Situational, personal, or temporal factors of using the mouse might shape the relationship between mouse usage and affect. For example, gender might be an important personal context variable given research showing gender differences in motor activity and emotional processing [56], [57]. Habituation to the mouse task or to the affect measurement might be factors in our study (see Figure 13 in Supplement 2 & 3 for a glimpse on some exploratory results on habituation). Other contextual factors include caffeine usage and the time of day at which the data were gathered, among many others. We largely omitted context variables in our data analysis to focus on the core bivariate relationship between mouse usage and affect. Incorporating these variables would have added excessive complexity. Yet, future studies should take these context factors into account for a more nuanced understanding of the relationship between mouse usage and affect.

7.1 Limitations

When discussing the results of this study, it is important to acknowledge that we used self-reported valence and arousal as the ground truth of affect. Self-report, while widely used as the ground truth in affective computing [58], is an imperfect measure of affect as it relies on self-awareness, subjective judgment, and belief [32]. Moreover, since subjective experience is just one component of affect [59], self-report, physiological and behavioral measures of affect may not necessarily correlate [60]. Future studies should therefore carefully select their ground truth measure of affect and consider multiple options.

Using EMA allowed to assess mouse usage together with affect in a naturalistic setting. However, the study design also comes with limitations [33, 34]. First, the provided valence and arousal ratings may not fully represent participants' affect if certain emotional states led them to skip data collection instances. Similarly, specific affective states might occur systematically less likely when the measurements take place. Consequently, the missing data points may not be randomly distributed. There was an observable within-person variance in the affect ratings. However, the overall distribution of valence and arousal were skewed towards a more positive and calmer affect. A lack of variance limits the possibility of accurate affect prediction. Second, while the field setting of this study is a strength in terms of external validity, it sacrifices control over participants. in possibly rare cases, multiple individuals might have used the same computer. Third, the study design does not permit testing for systematic changes in mouse usage in response to specific affective events. Instead, this study focuses on evaluating the relationship between mouse usage and ambient affect.

Our exhaustive use of the statistical toolbox to analyze the data can be considered a strongpoint, but there is no guarantee that we chose the best preprocessing options and analysis methods. Moreover, the data analysis followed a one-size-fits-all approach with identical model specifications for all mouse features, datasets and outcome variables. This was done to contain the complexity of this study. Future research could tailor model specifications to each mouse feature, dataset and outcome variable. For example, choosing a more sensitive feature selection procedure than simply removing highly correlated mouse features could decrease random noise in the data and improve prediction results.

Lastly, we consider the transformation of the raw mouse data into a specific set of mouse usage features a bottleneck when searching for a relationship between affect and mouse usage. A potentially infinite number of mouse features can be calculated from the raw data, thus any transformation into features entails loss of information. In the present study, we chose basic mouse features from previous studies and mouse data processing software. Recent studies [61] introduced advanced mouse usage features that might be more reliable predictors of emotion, especially with contextless mouse use.

The data analysis procedure in the present study underscored the importance of open science principles [62], [63], [64]. The data preprocessing and statistical modelling demonstrated that data analysis is a 'garden of forking paths [65]. The numerous options and decisions lead to a multiple comparison problem, which complicates the distinction between genuine evidence and supposedly meaningful noise [43]. Lastly, it is important to note that our data analysis and our interpretation of the results contain a degree of subjectivity. As the present results align with previous work of our research group [30], [31], one might be inclined to think that we conducted this study with a skeptical narrative in mind and analyzed as well as interpreted the results in an overly conservative way. Therefore, we encourage readers to carefully review the study material and data with a critical mind and draw their own conclusions.

8 Conclusions

The computer mouse offers an intriguing avenue for affect sensing in the field of affective computing due to its practicality. However, our study indicates that a definitive relationship between mouse usage and affect remains elusive. Future research is imperative to either uncover or conclusively dismiss the potential of the computer mouse for affect sensing.

Study Materials and Data

The source code of the Study-App is at https://doi.org/10.5281/zenodo.6559229. The data of the study are at https://doi.org/10.5281/zenodo.6559329. The analysis code of the study is at https://doi.org/10.5281/zenodo.10207296.

ACKNOWLEDGEMENT

We thank Johannes Blum, Christoph Rockstroh and Ranjit Singh for their valuable input. We thank Christina Häge for recruiting some participants.

REFERENCES

[1] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of Biomedical Informatics*, vol. 59, pp. 49–75, Feb. 2016, doi: 10.1016/j.jbi.2015.11.007

[2] J. B. Freeman, "Doing psychological science by hand," *Current Directions in Psychological Science*, vol. 27, no. 5, pp. 315–323, Aug. 2018, doi: 10.1177/0963721417746793

[3] D. U. Wulff, P. J. Kieslich, F. Henninger, J. M. B. Haslbeck, and M. Schulte-Mecklenbeck, "Movement tracking of cognitive processes: A tutorial using mousetrap," *PsyArXiv*, Dec. 2021, doi: 10.31234/osf.io/v685r

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan. 2010, doi: 10.1109/t-affc.2010.1

[5] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Trans. Affective Computing*, vol. 13, no. 1, pp. 440–460, Jul. 201, doi: 10.1109/taffc.2019.2927337

[6] P. Zimmermann, S. Guttormsen, B. Danuser, and P. Gomez, "Affective computing - A rationale for measuring mood with mouse and keyboard," *International Journal of Occupational Safety and Ergonomics*, vol. 9, no. 4, pp. 539–551, Jan. 2003, doi: 10.1080/10803548.2003.11076589

[7] D. Elliott, S. Hansen, L. E. Grierson, J. Lyons, S. J. Bennett, and S. J. Hayes, "Goal-directed aiming: Two components but multiple processes," *Psychological Bulletin*, vol. 136, no. 6, pp. 1023–1044, 2010, doi: 10.1037/a0020958

[8] E. Fox, "Perspectives from affective science on understanding the nature of emotion," *Brain and Neuroscience Advances*, vol. 2, Jan. 2018, doi: 10.1177/2398212818812628

[9] J. P. Gallivan, C. S. Chapman, D. M. Wolpert, and J. R. Flanagan, "Decision-making in sensorimotor control," *Nature Reviews Neuroscience*, vol. 19, no. 9, pp. 519–534, Sep. 2018, doi: 10.1038/s41583-018-0045-9

[10] J. A. Á. Martín, H. Gollee, J. Müller, and R. Murray-Smith, "Intermittent control as a model of mousemovements," *ACM Transactions on Computer-Human Interaction* (TOCHI), vol. 28, no. 5, pp. 1–46. Oct. 2021, doi: 10.1145/3461836

[11] G. F. Beatty and C. M. Janelle, "Emotion regulation and motor performance: An integrated review and proposal of the Temporal Influence Model of Emotion Regulation (TIMER)," *International Review of Sport and Exercise Psychology*, vol. 13, no. 1, pp. 266–296, 2020, doi: 10.1080/1750984x.2019.1695140

[12] J. Domínguez-Borràs and P. Vuilleumier, "Cognitive-Emotion Interactions," in The Cambridge Handbook of Human Affective Neuroscience, J. Armony and P. Vuilleumier, Eds. Cambridge: Cambridge University Press, 2013, pp. 330–416

[13] M. W. Eysenck, N. Derakshan, R. Santos, and M. G. Calvo, "Anxiety and cognitive performance: Attentional control theory," *Emotion*, vol. 7, no. 2, pp. 336–353, May 2007, doi: 10.1037/1528-3542.7.2.336

[14] J. P. Gallivan, C. S. Chapman, D. M. Wolpert, and J. R. Flanagan, "Decision-making in sensorimotor control," *Nature Reviews Neuroscience*, vol. 19, no. 9, pp. 519–534, Sep. 2018, doi: 10.1038/s41583-018-0045-9

[15] A. M. Mattek, P. J. Whalen, J. L. Berkowitz, and J. B. Freeman, "Differential effects of cognitive load on subjective versus motor responses to ambiguously valenced facial expressions," *Emotion*, vol. 16, no. 6, pp. 929–936, 2016, doi: 10.1037/emo0000148

[16] A. Nieuwenhuys and R. R. D. Oudejans, "Anxiety and perceptual-motor performance: toward an integrated model of concepts, mechanisms, and processes," Psychol. Res., vol. 76, no. 6, pp. 747–759, 2012, doi: 10.1007/s00426-011-0384-x

[17] C. M. Coelho, O. V. Lipp, W. Marinovic, G. Wallis, and S. Riek, "Increased corticospinal excitability induced by unpleasant visual stimuli," *Neurosci. Lett.*, vol. 481, no. 3, pp. 135–138, 2010, doi: 10.1016/j.neulet.2010.03.027

[18] B. Laursen, B. R. Jensen, A. H. Garde, and A. H. Jørgensen, "Effect of mental and physical demands on muscular activity during the use of a computer mouse and a keyboard," *Scand. J. Work Environ. Health*, vol. 28, no. 4, pp. 215–221, 2002, doi: 10.5271/sjweh.668

[19] G. F. Beatty, N. M. Cranley, G. Carnaby, and C. M. Janelle, "Emotions predictably modify response times in the initiation of human motor actions: A meta-analytic review," *Emotion*, vol. 16, no. 2, pp. 237–251, 2016, doi: 10.1037/emo0000115

[20] S. A. Coombes, K. M. Gamble, J. H. Cauraugh, and C. M. Janelle, "Emotional states alter force control during a feedback occluded motor task," *Emotion*, vol. 8, no. 1, pp. 104–113, 2008, doi: 10.1037/1528-3542.8.1.104

[21] A. J. Elliot and H. Aarts, "Perception of the color red enhances the force and velocity of motor output," *Emotion*, vol. 11, no. 2, pp. 445–449, Apr. 2011, doi: 10.1037/a0022599

[22] Y. Lu, K. J. Jaquess, B. D. Hatfield, C. Zhou, and H. Li, "Valence and arousal of emotional stimuli impact cognitive-motor performance in an oddball task," *Biol. Psychol.*, vol. 125, pp. 105–114, 2017, doi: 10.1016/j.biopsycho.2017.02.010

[23] K. M. Naugle, S. A. Coombes, J. H. Cauraugh, and C. M. Janelle, "Influence of emotion on the control of low-level force production," Res. *Q. Exerc. Sport*, vol. 83, no. 2, pp. 353–358, 2012, doi: 10.1080/02701367.2012.10599867

[24] T. Kowatsch, F. Wahle, and A. Filler, "Design and lab experiment of a stress detection service based on mouse movements," in The 11th Mediterranean Conference on Information Systems (MCIS), 2017, pp. 1–17, doi: 2010.3929/ethz-b-000218580

[25] T. Yamauchi and K. Xiao, "Reading emotion from mouse cursor motions: Affective computing approach," *Cognitive Science*, vol. 42, no. 3, pp. 771–819, Nov. 2018, doi: 10.1111/cogs.12557

[26] N. Banholzer, S. Feuerriegel, E. Fleisch, G. F. Bauer, and T. Kowatsch, "Computer mouse movements as an indicator of work stress: Longitudinal observational field study," *Journal of Medical Internet Research*, vol. 23, no. 4, Apr. 2021, Art. no. e27121, doi: 10.2196/27121

[27] M. T. Hibbeln, J. L. Jenkins, C. Schneider, J. Valacich, and M. Weinmann, "How is your user feeling? Inferring emotion through human-computer interaction devices," *MIS Quarterly*, vol. 41, no. 1, pp. 1–21, Jan. 2017, doi: 10.25300/misq/2017/41.1.01

[28] L. Pepa, A. Sabatelli, L. Ciabattoni, A. Monteriù, F. Lamberti, and L. Morra, "Stress detection in computer users from keyboard and mouse dynamics," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 12–19, Feb. 2021, doi: 10.1109/tce.2020.3045228

[29] D. Sun, P. Paredes, and J. Canny, "MouStress: Detecting stress from mouse motion," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems – CHI*, Apr. 2014, pp. 61–70. doi: 10.1145/2556288.2557243

[30] P. Freihaut and A. S. Göritz, "Using the computer mouse for stress measurement - An empirical investigation and critical review," *International Journal of Human-Computer Studies*, vol. 145, Jan. 2021, Art. no. 102520, doi: 10.1016/j.ijhcs.2020.102520

[31] P. Freihaut, A. S. Göritz, C. Rockstroh, and J. Blum, "Tracking stress via the computer mouse? Promises and challenges of a potential behavioral stress marker," *Behavior Research Methods*, vol. 53, no. 6, pp. 2281–2301, Apr. 2021, doi: 10.3758/s13428-021-01568-8

[32] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019, doi: 10.1177/1529100619832930

[33] C. Wrzus, and A. B. Neubauer, „Ecological Momentary Assessment: A meta-analysis on designs, samples, and compliance across research fields," *Assessment*, vol. 30, no. 3, pp.825–846, 2023, doi: 10.1177/10731911211067538

[34] A. A. Stone, S. Schneider, and J. M. Smyth, "Evaluation of pressing issues in ecological momentary assessment," *Annual Review of Clinical Psychology*, vol. 19, pp. 107-131. May 2023, doi: 10.1146/annurev-clinpsy-080921-083128

[35] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological Review*, vol. 110, no. 1, pp. 145–172, Jan. 2003, doi: 10.1037/0033-295x.110.1.145

[36] C. Geiser, T. Götz, F. Preckel, and P. A. Freund, "States and traits: Theories, models, and assessment," *Eur. J. Psychol. Assess*., vol. 33, no. 4, pp. 219–223, 2017, doi: 10.1027/1015-5759/a000413

[37] L. Hoffman, Longitudinal Analysis. Routledge, 2015, doi: 10.4324/9781315744094

[38] A. S. Göritz, "Building and managing an online panel with phpPanelAdmin," *Behavioral Research Methods*, vol. 41, no. 4, pp. 1177–1182, Nov. 2009. doi: 10.3758/brm.41.4.1177

[39] A. S. Göritz, "Determinants of the starting rate and the completion rate in online panel studies," in *Online Panel Research: Data Quality Perspective*, M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnik, and P. J. Lavrakas, Eds. John Wiley & Sons, Inc, 2014, pp. 154–170, doi: 10.1002/9781118763520.ch7

[40] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994, doi: 10.1016/0005-7916(94)90063-9

[41] J. Song, E. Howe, J. R. Oltmanns,and A. J. Fisher, "Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, vol. 30, no. 5, pp. 1662-1671, 2023, doi: 10.1177/10731911221113563

[42] P. J. Kieslich and F. Henninger, "Mousetrap: An integrated, open-source mouse-tracking package," *Behavior Research Methods*, vol. 49, no. 5, pp. 1652–1667, Jun. 2017, doi: 10.3758/s13428-017-0900-z

[43] S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel, "Increasing transparency through a multiverse analysis," *Perspectives on Psychological Science*, vol. 11, no. 5, pp. 702–712, Sep. 2016, doi: 10.1177/1745691616658637

[44] R. K. Singh, "Harmonizing instruments with equating," *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, vol. 6, no. 1, pp. 168–68 262, Aug. 2020, pid: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-68262-1

[45] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nat. Methods*, vol. 15, pp. 233–234, Apr. 2018, doi: 10.1038/nmeth.4642

[46] A. Gelman and J. Hill, Data Analysis Using Regression and multilevel/Hierarchical Models. Cambridge University Press, 2006, doi: 10.1017/CBO9780511790942

[47] A. J. Bishara and J. B. Hittner, "Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches," *Psychol. Methods*, vol. 17, no. 3, pp. 399–417, 2012, doi: 10.1037/a0028087

[48] A. Fernández-Fontelo, P. J. Kieslich, F. Henninger, F. Kreuter, and S. Greven, "Predicting question difficulty in web surveys: A machine learning approach based on mouse movement features," *Social Science Computer Review*, Jul. 2021, doi: 10.1177/08944393211032950

[49] J. Bergstra and Y. Bengio, "Random search for hyperparameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 2, pp. 281–305, Aug. 2012, [Online]. Available: http://jmlr.org/papers/v13/bergstra12a.html

[50] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/a:1010933404324

[51] S. Nakagawa, P. C. D. Johnson, and H. Schielzeth, "The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded," *Journal of The Royal Society Interface*, vol. 14, no. 134, Sep. 2017, doi: 10.1098/rsif.2017.0213

[52] Y. Benjamini and Y. Hochburg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing" *Journal of The Royal Society Interface*, vol. 57, no. 1, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x

[53] G. J. Martinez et al., "Alignment between heart rate variability from fitness trackers and perceived stress: Perspectives from a large-scale in situ longitudinal study of Information Workers," JMIR Human Factors, vol. 9, no. 3, 2022. doi:10.2196/33754

[54] N. Banholzer. "Data and Code: Computer Mouse Movements as an Indicator of Work Stress: Longitudinal Observational Field Study" June 28, 2021. Distributed by OSF. doi: 10.17605/OSF.IO/HE3F2

[55] L. F. Barrett and B. L. Finlay, "Concepts, goals and the control of survival-related behaviors," *Curr. Opin. Behav. Sci*., vol. 24, pp. 172–179, 2018, doi: 10.1016/j.cobeha.2018.10.001

[56] M. E. Kret and B. De Gelder, "A review on sex differences in processing emotional signals," *Neuropsychologia*, vol. 50, no. 7, pp. 1211–1221, 2012, doi: 10.1016/j.neuropsychologia.2011.12.022

[57] T. Yamauchi, J. H. Seo, N. Jett, G. Parks, and C. Bowman, "Gender differences in mouse and cursor movements," *Int. J. Hum. Comput. Interact*., vol. 31, no. 12, pp. 911–921, 2015, doi: 10.1080/10447318.2015.1072787

[58] S. Petrovica, A. Anohina-Naumeca, and H. K. Ekenel, "Emotion recognition in affective tutoring systems: Collection of ground-truth data," *Procedia Computer Science*, vol. 104, pp. 437–444, 2017, doi: 10.1016/j.procs.2017.01.157

[59] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," in *Appraisal Processes in Emotion: Theory, Methods, Research*, K. R. Scherer, A. Schorr, and T. Johnstone, Eds. Oxford University Press, 2001, pp. 92–120.

[60] J. Dang, K. M. King, and M. Inzlicht, "Why are self-report and behavioral measures weakly correlated?" *Trends in Cognitive Sciences*, vol. 24, no. 4, pp. 267–269, Apr. 2020, doi: 10.1016/j.tics.2020.01.007

[61] M. Weinmann, J. Valacich, C. Schneider, J. Jenkins, and M. Hibbeln, "The path of the righteous: Using trace data to understand fraud decisions in real time," *MIS Quarterly*, vol. 46, no. 4, pp. 2317–2336, 2022. doi:10.25300/misq/2022/17038

[62] J. B. Asendorpf, M. Conner, F. De Fruyt, J. De Houwer, J. J. A. Denissen, K. Fiedler, S. Fiedler, D. C. Funder, R. Kliegl, B. A. Nosek, M. Perugini, B. W. Roberts, M. Schmitt, M. A. G. Van Aken, H. Weber, and J. M. Wicherts, "Recommendations for increasing replicability in psychology," *Eur. J. Pers*, vol. 27, no. 2, pp. 108–119, Mar. 2013, doi: 10.1002/per.1919

[63] T. K. Bauer, U. Ebner-Priemer, M. Eid, A. S. Göritz, C. Lange, K. Maaz, E. Nagel, B. Raum, D. Richter, and M. Trappmann, "Data collection using new information technology: Recommendations on data quality, data management, research ethics, and data protection," *RatSWD Output*, vol. 6, no. 6, 2020, doi: 10.17620/02671.51

[64] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, Aug. 2015, doi: 10.1126/science.aac4716

[65] A. Gelman and E. Loken, "The statistical crisis in science," *American Scientist*, vol. 102, no. 6, 2014, doi: 10.1511/2014.111.460

Paul Freihaut received a BSc degree in psychology from Heidelberg University, in 2015, and an MSc degree in psychology from Goethe University Frankfurt, in 2017. He worked as a research associate and PhD student at the department of Psychology at the University of Freiburg and the Chair of Behavioral Health Technology at the University of Augsburg. His research interests include the use of computer technology and sensors in psychological research. His ORCID is: https://orcid.org/0000-0001-6249-5940.

Anja S. Göritz (https://www.goeritz.net) is full professor and holds the Chair of Behavioral Health Technology at the University of Augsburg. She studied psychology in Leipzig and completed her doctorate at the University of Erlangen-Nuremberg in Germany. She has published in journals including Behavior Research Methods, Leadership Quarterly and Journal of Business Ethics. She holds and manages an online access panel https:/wisopanel.net for scientific research. Her ORCID is: https://orcid.org/0000-0002-4638-0489.