Feasibility of Longitudinal Eye-Gaze Tracking in the Workplace

STEPHEN HUTT, University of Pennsylvania, USA
ANGELA E.B. STEWART, Carnegie Mellon University, USA
JULIE GREGG, University of Colorado at Boulder, USA
STEPHEN MATTINGLY, University of Notre Dame, USA
SIDNEY K. D'MELLO, University of Colorado at Boulder, USA

Eye movements provide a window into cognitive processes, but much of the research harnessing this data has been confined to the laboratory. We address whether eye gaze can be passively, reliably, and privately recorded in real-world environments across extended timeframes using commercial-off-the-shelf (COTS) sensors. We recorded eye gaze data from a COTS tracker embedded in participants (N=20) work environments at pseudorandom intervals across a two-week period. We found that valid samples were recorded approximately 30% of the time despite calibrating the eye tracker only once and without placing any other restrictions on participants. The number of valid samples decreased over days with the degree of decrease dependent on contextual variables (i.e., frequency of video conferencing) and individual difference attributes (e.g., sleep quality and multitasking ability). Participants reported that sensors did not change or impact their work. Our findings suggest the potential for the collection of eye-gaze in authentic environments.

CCS Concepts • **Human-centered computing~Ubiquitous and mobile computing systems and tools** • Human-centered computing~User studies

Additional Key Words and Phrases: Eye gaze, workplace, longitudinal data collection

ACM Reference format:
Stephen Hutt, Angela E.B. Stewart, Julie Gregg, Stephen Mattingly, and Sidney K. D'Mello 2022. Feasibility of Longitudinal Eye-Gaze Tracking in the Workplace. *Proc. ACM Hum.-Comput. Interact,* 6, CSCW1, Article 148 (May 2022), 22 pages, https://doi.org/10.1145/3530889

1 INTRODUCTION

Sensors such as eye trackers, cameras, electrocardiography, and electroencephalography have been instrumental in advancing the understanding of human thought, emotion, and behavior. From Duchenne's classic images of facial expressions [20], Jung's use of electrodermal activity to measure unconscious processes [42], and Huey's development of eye tracking to study reading

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800007, and the National Science Foundation (NSF; SES 2030599; SES 1928612; DRL 1920510).

2573-0142/2022/May – 148 $15.00
https://doi.org/10.1145/3530889

[37], the psychological sciences have immensely benefitted from developments in sensing technologies.

In recent years there has been an increasing discussion of the importance of studying people in ecological environments to ascertain the extent to which laboratory discoveries generalize in the real world and to make new discoveries by studying people in their natural environments. For example, [80] discusses the importance of investigating attention in naturalistic environments [17, 95]. The authors argue that although there are some similarities between attention in the lab and the real world, there are also a number of key differences [9, 13, 27]. For example, [34] examined applications of classical attention theory both in the laboratory and naturalistic environments. In the laboratory, participants completed a cueing task [28] where an on-screen face would provide a gaze cue prior to a gaze target. The cue would indicate the location of the next target, and participants were monitored as to how well they picked up on the cues. In the real world condition, visual cues were provided by a human being instead. Although the authors found that they could measure social attention in both environments, they did not find any reliable indices that were consistent across the two tasks. Similarly, [27] demonstrated that a model of eye movements in visual search of a static image did not transfer to a 3D environment (e.g., looking for an item in a picture of a room vs. looking for the item in that room).

Unfortunately, most sensing technologies used in psychological research are expensive and require a controlled laboratory environment. Not only does this impact ecological validity, but it also limits access to well-funded research labs. For example, the EyeLink 1000 is a state-of-the-art eye tracker that samples eye movements at up to 2000 Hz. This technology has led to numerous insights in understanding attention [73], information processing [75], working memory [81], and many other fields, such as marketing [93] and education [46]. However, the EyeLink 1000 costs over 30,000 dollars and is most successful when the head is stabilized in a chin rest/head support. As such, its use has been predominantly limited to laboratory environments where participants can be isolated from distractions (such as their phones, co-workers, family members, etc.), and the environment is strictly controlled (e.g., the head is held in a fixed position and lighting controlled). This precludes the use of the sensor in more naturalistic environments except in rare cases where the sensor itself is embedded in the environment (e.g., when eye trackers are embedded in surgical training environments, such as a simulated operating room [89]).

Some sensors are more amenable than others to the real world, and, in recent years, steps have been taken to move pertinent research out of the lab and into the wild. In particular, more individuals use wearable sensors, such as smartwatches and fitness trackers, as they go about their daily routines. Researchers have taken advantage of these sensors to understand individuals in their natural environments. Some early efforts include projects such as NetHealth [72], Project Tesserae [57], TILES [62], StudentLife [92], CampusLife [83], and WorkSense [54]. These studies have predominantly used cell phones and wearable sensors (providing physiological data such as heart rate and movement) to measure mood and health-related behaviors, such as activity [59], sleep [1, 2, 26, 56] as well as health events such as falls [67, 84]. Although wearable sensors have successfully measured physiological arousal, such as increased heart rate during exercise [69], it is less clear if these sensors can measure psychological constructs composed of both arousal and valence [82]. Valence is not as clearly expressed via physiological signals [22, 68]. It is also unclear if physiology can identify subtle differences among cognitive-affective blends, such as frustration and confusion [14].

In contrast, eye gaze is ideally suited to measure cognitive and affective states [15, 73], given well-established links between eye gaze and cognition [46, 73, 88]. The recent availability of

consumer off-the-shelf (COTS) eye trackers (retailing for several hundred dollars) has ushered forth an exciting era by enabling scalable "in the wild" research and applications (e.g., [39, 40, 99]). However, whether these COTS eye trackers provide valid measurements in real-world contexts, especially over extended periods of time, is unknown.

In this work, we investigate the feasibility of using COTS eye trackers to collect data in an ecologically valid environment. We developed a methodology to record and transmit data in the wild, requiring minimal experimenter oversight. We evaluate this approach by recording data from 20 information workers in their own workplaces over two weeks. We focused on the workplace as this is an ecologically valid space with a high number of potential applications; however, it also comes with a number of challenges. These include varying work environments, differing hardware and software setups, and privacy concerns. In addition, eye tracking is susceptible to data loss as a result of changes in head position or movement. Whereas previous research in the wild has typically considered data collected over a limited time window in the presence of researchers [4, 12, 39], we investigate the potential of using COTS eye trackers for continuous, longitudinal measurements in the workplace without continual researcher oversight. Our findings will be relevant to research and applications that aim to leverage gaze tracking for psychological assessment in authentic environments.

1.1 Challenges and Considerations

There are a number of challenges to overcome when tracking eye movements in the workplace compared to a controlled laboratory environment or wearable physiological devices. These challenges increase when factoring in sensing for extended time periods (days vs. hours).

Varying Office Environments. There are several different office setups in any workplace, with some having their own office, others sharing offices, or an open-plan office space. Recently, more people are working from home, adding even more potential variation in environments. This presents the potential for increased variability in data quality.

Multiple Displays. A person's computer setup can also vary considerably in terms of hardware (number of monitors, monitor size, processor power, etc.) and software (e.g., operating system). In cases where there are multiple displays, eye tracking must either occur across displays (which current gaze trackers do not support), or there needs to be a mechanism to select an appropriate display to track eye gaze.

Resource Management. A further challenge is that tracking must not place an undue burden on computing devices. To observe participants regular work behaviors, long-term sensing should not interfere with said regular work. In order to avoid slowing down the person's computer, tracking should only use available processor time, which can be difficult to quantify. As such, resources must be used sparingly to minimize the risk of impacting the user's regular work.

Tracker Calibration. In a typical one-hour research study, a user may calibrate an eye tracker multiple times in order to ensure the highest quality data [25, 65]. Calibration typically entails asking the user to focus on a series of points on the screen, allowing software to derive parameters required to convert raw gaze vectors into pixel coordinates [33]. Repeated calibrations are likely infeasible in the workplace as they would be too disruptive to the user.

User Privacy. Research has consistently strived to protect the identity of participants, ensuring that users are fully aware of their rights through informed consent and using standard techniques such as deidentification (e.g., removing name, date of birth, etc.). However, it is sometimes the case that participants can be reidentified from the data [79].

Recent work has attempted to further protect users through differential privacy [51], which involves adding noise to a sensor stream in a pseudorandom manner (often using a Gaussian process). This additional noise helps to protect a user from being identified from their data at a later point but introduces a privacy-utility trade-off. The increase in noise has the potential to limit the conclusions that can be drawn from the data. However, the hope is that with a large enough sample, the noise becomes irrelevant to the overall patterns in the data while still protecting the individual.

A further concern is the perception of privacy by the user. If users feel they are being observed or evaluated in some way, they are likely to behave differently [30, 45], limiting the ecological validity. An additional privacy concern is that of a bystander. Real-world sensing must take steps to ensure that only the user is being tracked and not bystanders, for example, when a user and a co-worker are viewing the screen together.

1.2 Background and Related Work

Eye tracking has been used as a research tool for decades [37], with a long history of laboratory studies examining attention [97], reading [5], visual search [76], and human-computer interaction [41]. For example, eye tracking has been used to measure confusion [47], frustration [44], and even personality traits [8]. Eye gaze research has also moved out of the lab; for example, gaze-based interaction has been used for military training in flight simulations [94], for target identification [35], and to help surgeons critically analyze their surgical skills [3]. Although these applications were designed for use outside the lab, they use research-grade eye trackers that cost thousands of dollars, thereby limiting widespread scalability.

Fortunately, the recent availability COTS eye trackers (retailing for hundreds of dollars rather than thousands) has enabled eye gaze research to move out of the lab (e.g., [58, 64, 99]). Though typically sampling at lower rates than research-grade equipment (i.e., 90Hz vs. 1000Hz), these trackers provide affordable and portable eye tracking for a fraction of the cost of research-grade equipment. COTS trackers are also known to be less accurate than research-grade equipment; however, research has indicated that such eye trackers could yield valid measurements in real-world contexts, such as classrooms [39], but with the presence of trained researchers to address technical problems. These studies have also only considered short-term tasks with repeated calibration of gaze rather than longitudinal tracking where calibration frequency is more limited.

Cheaper still is the option to use data from traditional RGB webcams to produce gaze estimates. Traditional eye tracking requires specialized equipment in order to illuminate the eye with infrared light and record the reflection produced [31]. In contrast, appearance- [6, 7, 96] and shape- [74, 78] based techniques can be used to generate gaze estimations from RGB images captured by a webcam. Converting these measures to screen coordinates has been consistently shown to be imprecise [43, 50] and to require multiple calibrations [98]. For example, a recent study reported a mean error of 10 degrees [98] (lower for some mobile devices) for a camera-based eye tracker. To place this into context, the EyeLink 1000, a state-of-the-art research-grade eye tracker, reports an average error of 0.3 degrees and considers an error above 1.5 to be too low quality[1]. The Tobii 4C, the COTS eye tracker used in our study, reports an average error of 0.6 degrees [29]. A notable exception here is [90], which examined how the smartphone cameras could be used for eye tracking after a brief calibration process. This work demonstrated equivalence in accuracy (though not in sampling rate) to mobile PCCR trackers (Tobii Pro 2 glasses) across four tasks. However, this study was conducted in a controlled lab environment where participants could be instructed on how to sit/position themselves. The relatively low

sampling rate of standard video (often less than 30hz in comparison to upwards of 60hz from a PCCR tracker) limits the computation of several critical gaze features, such as saccade onsets and smooth pursuits [77]. The need for calibration prior to each use also restricts applicability as noted above. Thus, although it is possible to estimate eye-gaze from webcam-based recordings, the state-of-the-art approach is not a viable replacement for gaze trackers for modeling users' mental states.

In psychology, longitudinal eye tracking has been a tool for learning about participants over time, for example, skill development [52] and the progressions of neurological conditions [18, 71]. In most cases, these involve bringing participants to the lab for a series of visits, each involving a calibration process and completion of lab-based tasks. These studies demonstrate the potential value for eye tracking over extended periods of time, but are still limited by laboratory settings and repeated calibrations – limitations that this research aims to address.

1.3 Current Study

Eye movements can provide cues to peoples' psychological states, such as emotions and attention. Until recently, use of these technologies has been mostly limited to the laboratory due to cost, calibration demands, setup, and other logistical issues. The recent introduction of low-cost, portable eye trackers has the potential of finally breaking away from the lab into real-world sensing. However, the question of whether COTS eye gaze can yield valid data in real-world environments and over extended periods of time remains unanswered.

We address this question by investigating the feasibility of ubiquitous monitoring of users' eye gaze in a complex authentic environment – the workplace. We conducted a study in which eye gaze of 20 information workers was recorded at pseudorandom time intervals throughout the workday over a two-week period. Because gaze tracking with webcam data is largely inaccurate (as reviewed above), we chose to use a dedicated COTS eye tracker (Tobii 4C) to monitor eye gaze, calibrated only once at the onset of the study. To the best of our knowledge, our study is the first to investigate the feasibility of collecting longitudinal eye gaze data in an authentic work context with minimal researcher involvement.

We had five research questions (RQs): (RQ1) To what extent can valid gaze data be collected in an ecologically valid setting over a two-week period from a single calibration? (RQ2) Does the validity of data degrade over time? (RQ3) How is data validity impacted by individual differences in terms of user traits and user environments? (RQ4) How consistent are key gaze measures derived from this approach compared to the literature? (RQ5) How does long-term sensing influence participants' perceptions of privacy and/or their regular work patterns?

2 METHODS

2.1 Participants

We recruited 21 participants from a large study focused on using sensors (wearable fitness devices, phone agents, etc.) to analyze traits, mental states, health, and workplace performance (see Mattingly et al., 2019 for an overview of this study). The current study commenced after the previous study had ended. The study (both the broader study and this sub-study) received approval from the university's institutional review board.

All participants involved in the study worked on-campus at a private university in the Midwest US. Participants all reported primarily working in an office; either their own office [n=9], a shared office [n=2], or an open-plan office environment [n=10]. All participants used operating systems compatible with the study software (Windows 7 [n=8], 8 [n=3], or 10 [n=10]), as well as having

computer setups suitable for the eye tracker(i.e., desktops [n=2] or laptops with a secondary monitor [n=19]). Participants also reported how frequently they used videoconferencing using a four-point scale (Never [n=5], Rarely [n=6], Sometimes [n=9], and Very Often [n=1]). In addition to providing written or electronic informed consent to participate in the larger study, participants also consented to the recording eye gaze in the workplace for additional compensation of $50.

Participants had the eye tracker installed for approximately two weeks. However, due to scheduling constraints, the number of days between installation and deinstallation ranged from 10 to 32, with a mean of 17 days and a median of 14 days. Due to a computer failure resulting in extensive data loss, one user was excluded from the analyses. Thus, data from 20 participants was analyzed here.

2.2 Materials

2.2.1 Equipment

Eye movements were recorded using a Tobii 4C[2], a COTS eye-tracking unit sampling at a 30 to 90 Hz variable rate, which automatically adjusted based on system performance. We opted to use a researcher for setup to ensure that the devices were correctly configured so we could have a uniform basis to analyze data quality over time, which is our central research question. The extent to which users could set up the devices on their own is a different question that we do not consider here, as, in most workplaces, staff configures a user's workspace. Similarly, the researcher delivered eye trackers to the participants, anticipating that future applications of this work would also entail enrolling participants and assisting with the initial setup. The eye tracker was affixed (via magnets in most cases) below the bottom of the participant's screen. If the user had multiple screens, the eye tracker was placed below the primary screen and calibrated for that screen only. The researcher confirmed the firewall settings allowed communication with our remote server and tested the data transfer (discussed below).

The eye tracker was calibrated once during the initial setup using a nine-point calibration system from the Tobii API[3], where nine points appear on the screen in turn, and the participant shifts their eye gaze from point to point. Participants were also shown the region in which their eyes would be tracked and how much movement was possible.

When launched (see below), the customized gaze recording software used this initial calibration. For each sample, we recorded the participant ID, the exact timestamp (in Coordinated Universal Time) for synchronization purposes, screen size (in pixels), the gaze location of each eye, the pupil diameter of each eye, and the eye tracker's internal validity value. The entire recording process required no input from the user, nor was the user notified in any way. The only indication that gaze was being monitored was a faint light on the front of the eye tracker.

2.2.2 Software

After the initial setup, participants were asked to resume their normal activities, and the approximate two-week data recording period commenced. We developed customized software to facilitate data collection and storage.

Sampling Method. We chose to sample participants rather than continuously record them so as to reduce the volume of data and ensure that data could be effectively transferred to the cloud-based storage servers in a reasonable time frame. Sampling was performed via a background process initiated at login. Every five minutes, the background process would randomly sample from a uniform distribution and trigger a 5-minute recording session if the sampled value was

greater than a prespecified threshold (.19). The threshold was set to yield, on average, five sessions per user per day. As we did not yet know the exact impact of the eye tracking, this number was selected to balance two goals (1) collecting enough data for analysis and (2) not oversampling or potentially overtaxing the participant's computer. To further prevent oversampling, once a recording had been initiated, a further recording could not be initiated for 30 minutes[4]. Sampling was limited to weekdays between the hours of 9 am and 5 pm. This reduced the likelihood of participants being recorded while in their homes (for laptop users) or when they were not working.

Data Transfer. Each data recording session was uniquely named with a session ID consisting of the anonymized participant ID and a timestamp. We appended the session ID to a file at the end of the five-minute data recording period to maintain a list of completed sessions. In addition, we maintained a second list of sessions that had been transmitted to the remote server. Before transmitting data, we checked both lists. If a session was "completed" and not yet "transmitted," we initiated the data transmission.

Data was transmitted to a remote server using HTTPS posts in Python. Data was first converted from tabular format to a JSON object. Only one sample of data was transmitted at a time. An access token was embedded into the header of the HTTPS post of the JSON object to ensure secure access to the remote server. After all samples of data were sent to the server, we transmitted a summary of the sent data, which detailed the number of valid and invalid records. After data transmission was finished, the session ID was appended to the list of transmitted data files for internal data checks. As this was an initial experiment, a local copy of all data was stored on participants' computers as a backup.

Data transfer was initiated at the end of each recording session and was monitored by researchers throughout the recording period via a Slack channel, which displayed how many points of data had been transferred to the server in the last 24 hours for participant ID. If no data transfer was observed over a 1-2 day period, and this period did not coincide with a weekend or a previously disclosed vacation, a researcher contacted the participant to troubleshoot the problem. This was done to understand whether the source of missing data was technical (participant disconnecting either the sensors), personal (unplanned or undisclosed -to the researchers- absence from the office), or for some other reason. In the former case, participants were instructed to reconnect the devices, which restored their functioning. We opted to correct technical issues because our research question is not concerned with whether participants have the necessary expertise to run the sensors on their own, but instead, the number of valid samples when the sensors function as intended. Having support from information technology professionals to diagnose and correct technical problems is standard in most office settings, so this aspect does not reduce the ecological validity of our study.

2.2.3 Preserving User Privacy

Each user was assigned a random alphanumeric identifier to index all of their data. Further protection of user privacy has been shown through methods such as differential privacy [51]; however, this involves adding noise to data as it is recorded [51, 87]. As this was an initial feasibility study and we were unsure of the data quality, we decided against this procedure. As user perception is critical to effective privacy [63, 70], participants were shown a sample of collected data during installation to demonstrate that it would be impossible to recreate an image of their face from the data collected. We also informed the participants that none of the data

collected would be used by their employers to evaluate them, nor would it be shared outside the research team.

Bystander privacy presented more of a challenge. Because no video or images of participants were collected, it is impossible to infer if the data collected is from the participant or a bystander. The gaze tracking monitors the closest target, which we assume will be the participant since it was their workstation (see Section 4.3 for further discussion).

2.2.4 Follow-up Survey

We designed a short survey to learn more about the participants' experiences. The survey assessed participants' prior privacy views independent of the current setup [53], along with perceived privacy [53], usability [49], and distraction [66] with the present setup. There were three items per measure, and all items utilized a six-point Likert scale (shown to be more reliable than a five-point Likert scale [16, 19]). Each measure included at least one reverse-coded item.

Both prior privacy views and perceived privacy were assessed as it was important to be able to compare participants' baseline privacy views to how they felt during this study [53]. The prior privacy measure asked participants, in general, how much they agreed with three statements, such as *"It is very important to me that I am aware and knowledgeable about how my personal information will be used"*. In contrast, questions for the perceived privacy measure focused on perceptions of privacy pertaining to the study. Participants were first instructed to consider the time when they were involved with the study and then asked how much they agreed with three statements (e.g., *"I was concerned about threats to my personal privacy when additional sensors were running"* - this item is reverse coded).

As data were collected in the background (i.e., there was no user interface), usability questions were adapted from [49] and focused on the extent to which participants were able to complete their regular tasks. Again, participants were instructed to consider their time in the study and state their agreement with three statements (e.g., *"I was able to efficiently complete my work even when the additional sensors were running."*). For distraction [66], items focused on if the user found the additional sensors distracting (e.g., *"I was more distracted than usual because of the additional sensors"*). The usability and distraction measures provide insight into any burden or disruptions caused by the sensors and how much this hindered the participants' day-to-day work.

In addition, we asked participants about their regular working patterns, including if they typically worked from 9 am to 5 pm and the frequency at which they left their desks. We then asked participants, "Overall, how satisfied were you with your experience with the additional sensors" with responses again on a 6 point Likert scale. Finally, we asked participants two open-ended questions, one asking them to discuss any challenges completing their regular work because of the sensors, and the second requesting any additional comments.

The follow-up survey was administered approximately 30 weeks later. Ideally, the survey would have been completed immediately after deinstallation, but was delayed due to logistical issues pertaining to the main study. All 21 participants completed the survey, upon which they were compensated with an additional $10 gift card.

2.2.5 Trait Level Measures

Participants also individually completed a large battery of validated individual differences surveys as part of the larger study. Of these, we analyzed a subset in the current study to investigate whether these were related to the quality of data collection (RQ3), including the Morningness-Eveningness Questionnaire (MEQ) [36], the Pittsburgh Sleep Quality Index (PSQI)

[85], and SynWork [23]. These survey measures cover a wide range of constructs all related to the broader study on workplace performance. These measures provide insights on psychological traits (e.g., personality and affect) as well as physical characteristics (e.g., sleep) that relate to a users day to day activities. These surveys thus provide a convenient opportunity to explore how a users characteristics and traits relate to the eye tracking and identify any moderators that may need to be accounted for in long-term eye tracking.

The MEQ was originally developed by Horne and Ostberg [36] to determine whether a person's circadian rhythm (biological clock) produces peak alertness, either in the morning, in the evening, or in between. The survey consists of 19 multiple choice questions assessing a person's sleep habits. A higher score on the MEQ indicates peak alertness in the morning, whereas a lower score indicates peak alertness in the evening.

PSQI is a self-rated questionnaire that assesses sleep quality and disturbances [85]. Nineteen individual items generate seven "component" scores: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. The sum of these seven components provides one overall score. A lower PSQI score indicates healthier sleep patterns.

SYNWORK is a computer-based multitasking environment composed of synthetic work tasks that are relevant to a number of occupations first presented by Elsmore in 1994 [23]. SYNWORK is beneficial for experiments as it provides both experimental control and realism for the participants [21]. The task requires participants to engage in four individual tasks simultaneously, each focused on a different skill; Memory, Visual Monitoring, Auditory Monitoring, and Math. Participants used a mouse to complete each task. Incorrect task completion resulted in a sound, whereas there was no audible feedback for correct responses (as this tended to be distracting).

3 RESULTS

In total, there were 1,382 five-minute recording sessions across the 20 participants. Figure 1 shows a histogram of the number of sessions possible for each participant over the two-week period. We note that four of the participants had less than 45 potential sessions (times when the background triggered a recording session), ostensibly due to the computer being off or not being used during the time window. The median number of sessions per participant was 69 (mean = 68, SD = 31).

3.1 Number of Valid Samples (RQ1)

A sample was considered valid if at least one eye was accurately tracked, determined by the Tobii API. We deemed one eye tracked to be sufficient compared to a more stringent two eyes tracked criterion as features such as fixations and saccade approximations can be derived from one eye. Indeed it is common to only use data from one eye for these calculations, rather than merging data from two eyes [86].

For each gaze sample, the Tobii API provides a Boolean value for each eye, indicating if the eye was accurately detected. There are a number of reasons for an invalid sample. For example, the user could be looking away from the screen (e.g., down at the keyboard), could have closed their eyes, or may not currently be at their desk. Similarly, an invalid sample could indicate a technical issue, where the eyes could not be detected due to a calibration error, incorrect angling of the eye tracker, or interference from external light sources (both visible and infra-red).

Of the 1,382 sessions, 919 sessions (66.5%) had at least one sample of valid gaze data. All twenty participants were able to record some amount of gaze data with varying levels of success. Mean

session- and participant-level proportion of valid gaze samples was 30% and 32%, respectively, see Table 1 and Figure 1.

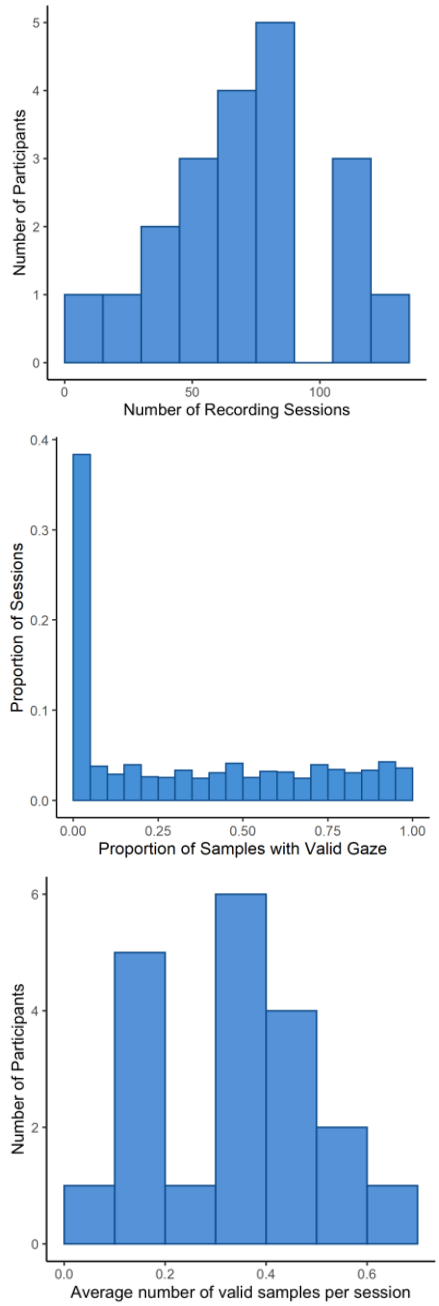

Figure 1. Histograms showing of the number of possible sessions per participant (top), the proportion of each session that had a valid gaze sample (middle), and the participant level average number of valid samples per session (bottom)

We also noted that the tracker occasionally would pick up eye gaze outside the screen bounds. If we limit gaze samples to those within the screen bounds, the average number of samples slightly decreased to 28% (session-level) and 29% (participant level). The distribution of valid samples when averaged at the participant level was significantly different than zero, $t(19)=8.76$, $p<0.001$. We deemed the number of valid samples recorded to be acceptable in that participants only calibrated their eye tracker once. Research has indicated that eye tracker calibration may drift over time, causing less and less valid samples to be collected [32, 33]. In most eye tracking studies, participants calibrate the tracker at least once per session, if not multiple times, which we did not do to preserve ecological validity (see above). Similarly, we have no control over whether a user is using their computer (or even at their desk) at the time of sampling, which is also likely to reduce the overall number of samples.

Table 1. Descriptive statistics for validity measures at session-level and participant-level

	Gaze (One Eye Tracked)		**Gaze (Both Eyes Tracked)**	
No. of sessions with at least one valid sample	919		892	
No. of sessions with no valid samples	463		490	
Proportion of Valid Samples	Session Level	Participant Level	Session Level	Participant Level
Mean	30%	32%	26%	28%
SD	33%	17%	31%	16%
Min	0%	2%	0%	1%
Max	100%	60%	99%	57%

3.2 Valid Samples Over Time and Contextual Influences (RQ2)

We examined how the number of valid samples varied over time. Each participant had the tracker for approximately two weeks, with only one calibration at the beginning of that period. To systematically examine the influence of time, we regressed the count of valid samples per session on session number (which indexes days from the start of the study) and time of the recording (which indexes hours in the day) using linear mixed effects models with participant as the random intercept due to the nested nature of the data (multiple sessions per participant). We found that number of valid samples decreased across sessions (B=-1.68, p= .0.057), suggesting that depending on the length of the tracking period, additional calibration may (at some point) be necessary. More work is needed to find the exact inflection point at which recalibration would be recommended. There was no effect of time of day (B = -.45, *p=.585*).

Next, we investigated whether contextual variables moderated above the influence of session number. Using linear mixed effects models, we included the following contextual variables as predictors and examined the time × context interaction term: (1) open/shared office [N=11] or own office [N=9]); (2) computer use (use of computers for work only [N = 14] or work and personal use [N = 6]); and (3) video conferencing (whether participants' never or rarely [N = 11] vs. sometimes or often [N = 9] participated in video conferencing). There were no significant interactions between session number and time of day with office layout (*ps* > .11) and computer use, but we observed a significant interaction effect of session number for those who often or regularly engaged in video conferencing (B = 3.90; p = 0.03). We probed this interaction using a simple slopes analysis which entails assessing the relationship between session number and valid

samples for different levels of the moderator (video conferencing). The results, shown in Figure 2, indicated that the number of valid gaze samples decreased over time for participants who rarely took video calls (B = -3.48; p = .01), this was not the case for participants that took video calls more often (B = .42, p = .73). We theorize that this may be due to those that rarely took video calls being less conscious of their position relative to the screen.

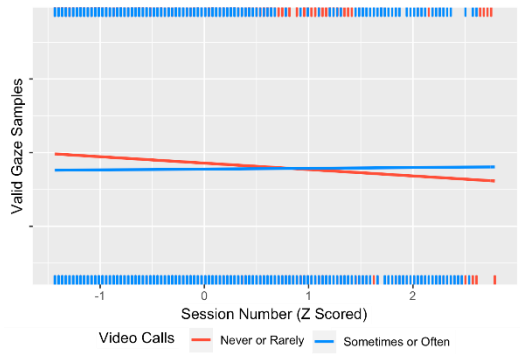

Figure 2. Interactions between frequency of video calls and the number of valid gaze samples

Next, we investigated the impact of trait level measures (SynWork score, PSQI score, and MEQ Score – see section 2.2.5). Trait level data was not available for one of the participants, so this analysis was completed with the remaining 19 participants. We again used mixed effects models as above and examined the time × trait interaction term. We observed no significant effect (main or interaction) with MEQ score, however we observed a significant interaction between session number and SynWork (B = 3.1, p < 0.01) and PSQI (B = -2.46, p = 0.01) scores. We probed both interactions again with simple slopes analyses (see Figure 3). We observed that participants with SynWork scores one standard deviation below the mean (low multitasking) were more likely to see degradation in the number of valid samples over sessions (B = -4.14, p < 0.01); whereas there the number of valid samples remained constant for average multitaskers (B = 0.92, p = .31) or even tended to increase (albeit nonsignificant, B = 1.39, p = .17) for good multitaskers (SynWork scores 1SD above the mean). Similarly, those with PSQI scores one SD above the mean (indicating less healthy sleep) were also more likely to have the number of valid samples degrade over time (B=-3.82, p < 0.01), whereas those with average PSQI scores showed a more consistent number of valid samples over time (B=-1.36 p=.14) and those with PSQI scores 1 SD below the mean (good sleep health) saw a slight (nonsignificant) increase in number of valid samples (B=1.11, p=.40). We also observed a significant interaction between PSQI and the hour of the recording (B=1.85, p=0.03). Simple slopes analysis indicated that participants with PSQI scores one SD below the mean (healthier sleep) were less likely to have a higher number of samples later in the day (B=-2.34, p=0.05), whereas those with PSQI scores above the mean (less healthy sleep) were more likely to have more valid gaze samples later in the day (B=1.35, p=.26).

3.3 Feature Extraction (RQ4)

We next examined if the data collected was of suitable fidelity for additional analysis. We first fixation filtered the raw eye gaze data using Open Gaze and Mouse Analyzer (OGAMA) [91]. Decades of eye tracking research have yielded a wide variety of potential features derived from fixations and fixation patterns [24, 38, 39, 60]. Given that we do not know what the user was doing

at the time of the gaze recording, the most relevant for this work was to examine the average fixation duration and compare to previously collected standards. The average fixation duration across all sessions ranged from 109ms to 760ms, with a mean of 373ms (see Figure 4). This finding is consistent with work on eye movements which suggests in most cases fixations can last between 50-600ms depending on the context [48, 76, 77]. Without additional data of screen context (that would have potential privacy implications), it is difficult to evaluate if these fixations were in fact accurate. However, this result reinforces the potential feasibility of our longitudinal data collection as derived features fall within expected ranges.

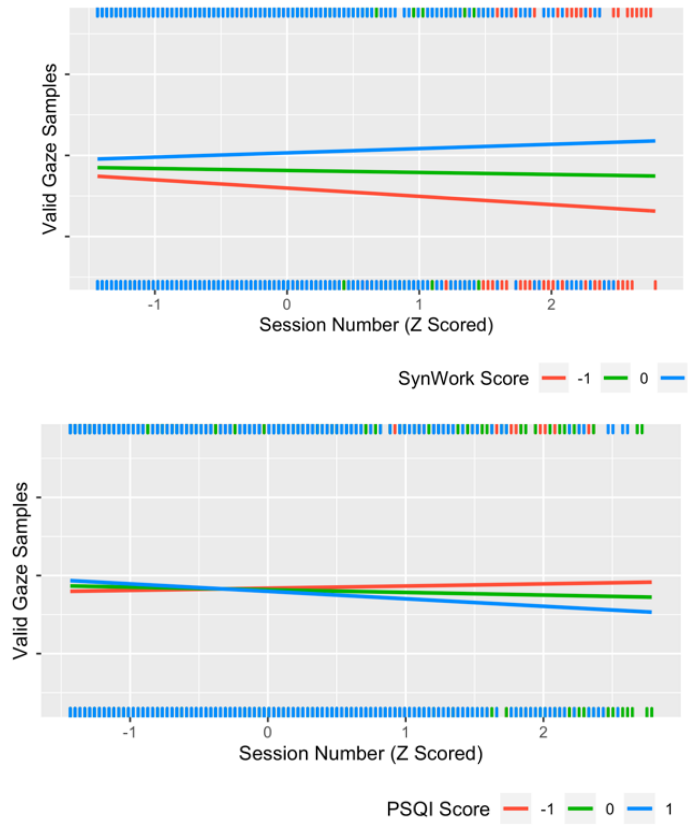

Figure 3 Interactions between session number and SynWork score z scored, (top) and PSQI Score z scored (bottom) when predicting number of valid samples in a session

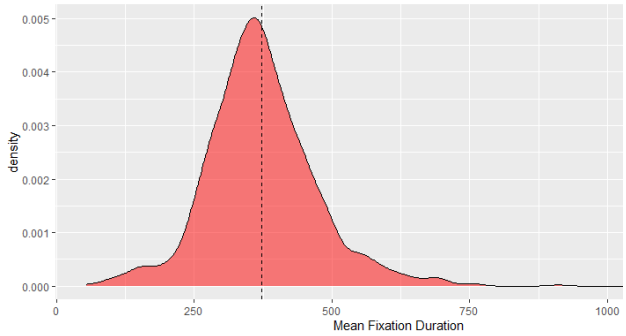

Figure 4. Density Plot of Average Fixation Duration per Session

3.4 User Perceptions (RQ5)

We analyzed the user survey to examine how participants perceived the eye tracker when incorporated into their work environments. Descriptive statistics for each of the individual measures are shown in Table 2. Participants reported a mean overall satisfaction of 5.14 (out of 6), suggesting that participants were generally satisfied with their experience. We also observed high (M = 5.11) usability scores, which in our case indicated the tracker did not interfere with participants' work (e.g., computer slowing down due to the recording process). Similarly, distraction scores were low (M = 2.06), suggesting that work routines were not interrupted by the eye tracker being distracting.

We used measures of prior privacy views to understand whether participants were sensitive to privacy issues, specifically regarding technology. We note that our participants would describe themselves as being somewhat sensitive to privacy issues (M = 4.43, a higher score indicating greater privacy concern). Despite this, perceived privacy during the study (a higher score indicating a greater sense of privacy) was also high (M = 4.89). This implies that despite being privacy conscious, participants felt a high level of privacy during the study.

We also analyzed participants' responses to the open-ended item - "Please describe any challenges you experienced with completing your usual work due to the additional sensors. Please be as explicit as possible." A researcher coded each of their responses. We note that 17/20 participants reported no challenges or interference to their work, and one user reported challenges with initial setup (the user that was excluded from analysis). The remaining responses are listed below and suggest some initial concerns with being monitored, occasional distraction from the eye tracker infrared lights, and minor annoyances with the sensors and software:

- "No real challenges. I did feel more watched, especially in the beginning, and there were a few times that the sensors needed upgrading/weren't connecting accurately and it took me a bit to fix that."
- "Occasionally the lights [on the eye tracker] turning on would be a distraction, but once they stayed on I forgot about them."
- "Occasionally, [the eye tracker] on the bottom of my monitor would fall off, but I just had to press it back on. Other than that, I didn't really notice or was impacted by the sensors."

Table 2. Descriptive Statistics for user survey measures, for each measure theoretical min and max, is 1 and 6 respectively.

Measure	Mean	S.D.	Min	Max
Overall Satisfaction	5.14	0.65	4.00	6.00
Prior Privacy Concern	4.43	0.76	2.33	6.00
Perceived Privacy	4.89	0.64	3.67	6.00
Usability	5.11	0.63	4.33	6.00
Distraction	2.06	0.77	1.00	3.33

4 DISCUSSION

Over the past decades, eye tracking has provided countless insights into psychological science [73]. Unfortunately, research-grade equipment is expensive and requires a controlled laboratory environment. Not only does this limit the ecological validity of findings, it also limits the research to those with the required resources. The development of newer, low cost, portable eye tracking, has opened the door to a number of potential research activities, particularly studying affect and attention "in-the-wild." However, it is not yet known how well COTS tracking will perform outside the controlled environment of the lab and what potential there is for longer-term sensing (e.g., not a single 1-hour study). To address these questions, we recorded eye movements from 20 users in real-world environments using COTS sensors. We investigate the validity of data collected and the effect of time and computer setup on recording quality. We also used surveys to assess user perceptions of the sensing setup. Our main findings are summarized below, followed by a discussion of applications, limitations, and future work.

4.1 Main Findings

Despite the noisy workplace environment, we found that it is feasible to collect valid gaze samples with COTS eye tracking in normal working environments and without any restrictions on the participants (RQ1). Researchers were not present for data collection (other than at installation), nor did they enforce any protocol/compliance from participants (beyond occasional check-ins). We achieved an average proportion of valid samples of 32%, which we consider moderate given the lack of constraints. For example, we could not control if the participants were using their computer or even at their desks. Even when working on their computers, participants could have turned to discuss something with a colleague or reached to get something from across the desk, all factors that would result in invalid signals. Further, although the proportion of valid samples may be lower than equivalent laboratory studies [10, 11], they are in line with other short-term (60-90 minute) "in the wild" studies with these sensors [11, 39].

We found that number of valid samples diminished over time (RQ2) with marginal significance. This implies that some level of recalibration is required, though how frequent is still an open question. We gained a deeper insight into signal degradation by examining how contextual and trait measures moderated the effect of time (RQ3). We observed significant degradation in the number of valid gaze samples collected per session over time for those that rarely took video calls – potentially because they were less conscious of their positioning. We also saw that participants with more unhealthy sleep habits or worse multitasking skills also had greater degradation over time. We did not observe any impact of morningness-eveningness (MEQ score) on the number of valid samples. These moderators give insights that will be critical as we design more longitudinal data collection.

We have shown that the data collected from the sensors is of suitable fidelity for feature calculation (RQ4), with calculated values with expected ranges based on previous findings. Finally, We found that participants positively perceived their experiences with the tracker (RQ5), reporting a mean overall satisfaction of 5.14 out of 6. Our participants also did not raise any privacy concerns with the tracking, even though they described themselves as generally concerned with privacy matters. Participants also found the long-term sensing to provide minimal distractors to their regular work. Taken together, our result suggests that longitudinal (at least within the two weeks considered here) gaze tracking is a viable option in the workplace despite the challenges outlined above.

4.2 Applications

Though this work has several applications, it is important first to discuss how this approach should *not* be used. It should never be used for "big brother" style monitoring or for evaluative assessments of workplace satisfaction and performance. It should mainly be used for research studies and workflow improvement efforts with full opt-in and consent of individual users, never mandated by supervisors or employers. Similarly, we should be conscious of the potential effects of adding this type of monitoring on participants health (e.g., if they feel they are being observed, are they less likely to take breaks, or move to stretch, etc.).

The results presented here pave the way to a wide variety of ecologically valid studies, such as those that explore multitasking, stress, and burnout in the workplace. Further, studies could, consider other ecologically valid environments such as remote schooling or work from home (e.g., Mark et al., 2016). Researchers could also examine if results obtained from laboratory experiments replicate in ecologically valid environments, and if they do not, analyze the differences.

In addition to psychological/cognitive studies, these results also present opportunities for usability studies. For example, developers can leverage eye gaze to better understand how a user is interacting with their software and identify general usage patterns as well as 'problem spots'. Understanding how users interact and respond to errors in their own environments will provide more ecological valid user studies that in turn could lead to better software interfaces.

More broadly, the success of longitudinal gaze tracking in the wild has several applications beyond research. In educational contexts, gaze tracking has been used to monitor affect and engagement and deliver automated interventions [39, 61]. Similar applications, such as an interactive assistant that delivers real-time data-driven feedback or suggestions, can be developed in workplace contexts. For example, if sensors indicate that the user is frustrated or cognitively overloaded, the assistant might suggest taking a break or temporarily switching to another task.

4.3 Limitations and Future Work

Like all studies, ours has limitations. First, our study was conducted with a small number of participants using a convenience sample of participants who were already part of a larger study monitoring them in the workplace [57]; these participants might be more predisposed to being monitored and less likely to have privacy concerns. Further, researchers monitored the data collection remotely and provided troubleshooting information if no recording sessions occurred, which may not always be possible for real-world use. We also considered just one type of eye tracker (the Tobi 4C). Future work should consider more diverse participants, multiple eye trackers, and collect more expansive measures of individual attributes (such as if a user wears eyeglasses).

Second, the number of sessions varied greatly between participants. Analyzing the factors that caused variation across participants (e.g., did they typically work in the sample window) may lead to more accurate tracking in the future. Furthermore, some participants had 100% sample validity for some sessions. A deeper analysis of these sessions could potentially yield insights into how to improve tracking.

Future work should also further expand on the duration of the study. We observed a decrease in sample validity, implying that some level of repositioning or recalibration may be required for long-term tracking. Future work should attempt to pinpoint how long a participant can record data before the need to recalibrate. It should also consider the effect that recalibration may have on user behavior and their perceptions of being monitored.

Although we have taken considerable steps to preserve user privacy, participants still have potential to be identified from their eye gaze [79] if a suitable comparison dataset is available. In recent work, methods have been proposed to preserve user privacy against such comparisons by systematically introducing noise into the dataset [87]. Future work should examine the impact of these differential privacy techniques on data quality. Similarly, we have not considered bystander privacy in this work. If two people were at the same workstation (e.g., co-editing a report), the bystanders' eye movements could have been recorded. Future work should examine the possibility of using face recognition or similar to guarantee that only the desired participant's data is collected.

4.4 Concluding Remarks

The recent introduction of commercial off-the-shelf eye-trackers has brought with it exciting opportunities to move research previously confined to the lab into more ecologically valid environments. This work has made three contributions toward this goal. First, we have shown that it is possible to collect gaze data with minimal experimenter oversight (other than the initial setup and technical troubleshooting) over a two-week period. Second, we have demonstrated that this method is somewhat robust over time, an essential consideration for long-term sensing. Third, we investigated the effect of the sensors on users' perceptions of privacy and on their workflow. On average, users reported little to no distraction from the sensors and minimal interruptions to their regular working habits, indicating that this method would be suitable for future studies striving for ecological validity. We also show that even though our users characterize themselves as privacy conscious, they feel comfortable with the data being collected by the sensors. Our findings suggest that it may finally be possible to apply decades of lab-based eye tracking research in the noisy workplace environment, thereby affording new discoveries regarding human behavior.

ACKNOWLEDGMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800007, and the National Science Foundation (NSF; SES 2030599; SES 1928612; DRL 1920510). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein.

5 REFERENCES

[1] Abdullah, S., Matthews, M., Murnane, E.L., Gay, G. and Choudhury, T. 2014. Towards circadian computing: "Early to bed and early to rise" makes some of us unhealthy and sleep deprived. *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014).https://doi.org/10.1145/2632048.2632100.

[2] Abdullah, S., Murnane, E.L., Matthews, M. and Choudhury, T. 2017. Circadian Computing: Sensing, Modeling, and Maintaining Biological Rhythms. *Mobile Health: Sensors, Analytic Methods, and Applications*. (2017).https://doi.org/10.1007/978-3-319-51394-2_3.

[3] Ahmidi, N., Hager, G.D., Ishii, L., Fichtinger, G., Gallia, G.L. and Ishii, M. 2010. Surgical Task and Skill Classification from Eye Tracking and Tool Motion in Minimally Invasive Surgery. *13th International Conference on Medical Image Computing and Computer-Assisted Intervention* (Beijing, China, Sep. 2010), 295–302.https://doi.org/10.1007/978-3-642-15711-0_37.

[4] Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K. and Christopherson, R. 2009. Emotion sensors go to school. *Frontiers in Artificial Intelligence and Applications* (2009), 17–24.https://doi.org/10.3233/978-1-60750-028-5-17.

[5] Asteriadis, S., Tzouveli, P., Karpouzis, K. and Kollias, S. 2009. Estimation of Behavioral User State Based on Eye Gaze and Head Pose-Application in an e-Learning Environment. *Multimedia Tools and Applications*. (2009).https://doi.org/10.1007/s11042-008-0240-1.

[6] Bacivarov, I., Ionita, M. and Corcoran, P. 2008. Statistical Models of Appearance for Eye Tracking and Eye-Blink Detection and Measurement. *IEEE Transactions on Consumer Electronics*. 54, 3 (2008), 1312–1328.https://doi.org/10.1109/TCE.2008.4637622.

[7] Baluja, S. and Pomerleau, D. 1994. Non-Intrusive Gaze Tracking Using Artificial Neural Networks, CMU-CS-94-102. *Neural Networks*. (1994), 753–760.

[8] Berkovsky, S., Taib, R., Koprinska, I., Wang, E., Zeng, Y., Li, J. and Kleitman, S. 2019. Detecting Personality Traits Using Eye-Tracking Data. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), 221:1--221:12.https://doi.org/10.1145/3290605.3300451.

[9] Birmingham, E., Johnston, K.H.S. and Iarocci, G. 2017. Spontaneous Gaze Selection and Following during Naturalistic Social Interactions in School-Aged Children and Adolescents with Autism Spectrum Disorder. *Canadian Journal of Experimental Psychology*. (2017).https://doi.org/10.1037/cep0000131.

[10] Bixler, R. and D'Mello, S.K. 2016. Automatic Gaze-Based User-Independent Detection of Mind Wandering during Computerized Reading. *User Modeling and User-Adapted Interaction*. 26, 1 (2016), 33–68.https://doi.org/10.1007/s11257-015-9167-1.

[11] Bosch, N. and D'Mello, S.K. 2019. Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing*. PP, c (2019), 1–1.https://doi.org/10.1109/taffc.2019.2908837.

[12] Bosch, N., D'Mello, S.K., Ocumpaugh, J., Baker, R.S. and Shute, V. 2016. Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. *ACM Transactions on Interactive Intelligent Systems* (2016), 1–26.https://doi.org/10.1145/2946837.

[13] Brennan, A.A., Bruderer, A.J., Liu-Ambrose, T., Handy, T.C. and Enns, J.T. 2017. Lifespan Changes in Attention Revisited: Everyday Visual Search. *Canadian Journal of Experimental Psychology*. (2017).https://doi.org/10.1037/cep0000130.

[14] Calvo, R.A., D'Mello, S.K., Gratch, J.M. and Kappas, A. 2015. The Oxford Handbook of Affective Computing. *The Oxford Handbook of Affective Computing*. (2015).https://doi.org/10.1093/oxfordhb/9780199942237.001.0001.

[15] Chita-Tegmark, M. 2016. Social Attention in ASD: A Review and Meta-Analysis of Eye-Tracking Studies. *Research in Developmental Disabilities*. 48, (2016), 79–93.

[16] Chomeya, R. 2010. Quality of Psychology Test Between Likert Scale 5 and 6 Points. *Journal of Social Sciences*. 6, 3 (2010), 399–403.https://doi.org/10.3844/jssp.2010.399.403.

[17] Cohen, G. and Conway, M.A. 2007. Memory in the Real World. (2007).

[18] Crawford, T. 2015. The Disengagement of Visual Attention in Alzheimer's Disease: A Longitudinal Eye-Tracking Study. *Frontiers in Aging Neuroscience*. 7, (2015).https://doi.org/10.3389/fnagi.2015.00118.

[19] Cummins, R.A. 2000. Why We Should Not Use 5-Point Likert Scales: The Case for Subjective Quality of Life Measurement. (2000).

[20] Darwin, C. 1872. The Expression of the Emotions in Man and Animals. (1872).

[21] DiFonzo, N., Hantula, D.A. and Bordia, P. 1998. Microworlds for Experimental Research: Having Your (Control and Collection) Cake, and Realism Too. *Behavior Research Methods, Instruments, and Computers*. (1998).https://doi.org/10.3758/BF03200656.

[22] Ekman, P., Levenson, R.W. and Friesen, W. V. 1983. Autonomic Nervous System Activity Distinguishes among Emotions. *Science*. (1983).https://doi.org/10.1126/science.6612338.

[23] Elsmore, T.F. 1994. SYNWORK1: A PC-Based Tool for Assessment of Performance in a Simulated Work Environment. *Behavior Research Methods, Instruments, & Computers*. (1994).https://doi.org/10.3758/BF03204659.

[24] Faber, M., Bixler, R. and D'Mello, S.K. 2018. An Automated Behavioral Measure of Mind Wandering during Computerized Reading. *Behavior Research Methods*. 50, 1 (2018), 134–150.https://doi.org/10.3758/s13428-017-0857-y.

[25] Faber, M., Krasich, K., Bixler, R.E., Brockmole, J.R. and D'Mello, S.K. 2020. The Eye-Mind Wandering Link: Identifying Gaze Indices of Mind Wandering Across Tasks. *Journal of Experimental Psychology: Human Perception and Performance*. (2020).https://doi.org/10.1037/xhp0000743.

[26] Fei, J. and Pavlidis, I. 2010. Thermistor at a Distance: Unobtrusive Measurement of Breathing. *IEEE Transactions on Biomedical Engineering*. (2010).https://doi.org/10.1109/TBME.2009.2032415.
[27] Foulsham, T. and Kingstone, A. 2017. Are Fixations in Static Natural Scenes a Useful Predictor of Attention in the Real World? *Canadian Journal of Experimental Psychology*. (2017).https://doi.org/10.1037/cep0000125.
[28] Friesen, C.K. and Kingstone, A. 2003. Abrupt Onsets and Gaze Direction Cues Trigger Independent Reflexive Attentional Effects. *Cognition*. 87, 1 (2003), B1–B10.https://doi.org/https://doi.org/10.1016/S0010-0277(02)00181-6.
[29] Gibaldi, A., Vanegas, M., Bex, P.J. and Maiello, G. 2017. Evaluation of the Tobii EyeX Eye Tracking Controller and Matlab Toolkit for Research. *Behavior Research Methods*. (2017).https://doi.org/10.3758/s13428-016-0762-9.
[30] Goodman, E., Kuniavsky, M. and Moed, A. 2013. Observing the User Experience: A Practitioner's Guide to User Research (Second Edition). *IEEE Transactions on Professional Communication*. (2013).https://doi.org/10.1109/tpc.2013.2274110.
[31] Guestrin, E.D. and Eizenman, M. 2006. General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections. *IEEE Transactions on Biomedical Engineering*. (2006).https://doi.org/10.1109/TBME.2005.863952.
[32] Hansen, D.W. and Pece, A.E.C. 2005. Eye Tracking in the Wild. *Computer Vision and Image Understanding*. (2005).https://doi.org/10.1016/j.cviu.2004.07.013.
[33] Harezlak, K., Kasprowski, P. and Stasch, M. 2014. Towards accurate eye tracker calibration -methods and procedures. *Procedia Computer Science* (2014).https://doi.org/10.1016/j.procs.2014.08.194.
[34] Hayward, D.A., Voorhies, W., Morris, J.L., Capozzi, F. and Ristic, J. 2017. Staring Reality in the Face: A Comparison of Social Attention Across Laboratory and Real World Measures Suggests Little Common Ground. *Canadian Journal of Experimental Psychology*. (2017).https://doi.org/10.1037/cep0000117.
[35] Hild, J., Kühnle, C. and Beyerer, J. 2016. Gaze-based Moving Target Acquisition in Real-time Full Motion Video. *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (New York, NY, USA, 2016), 241–244.https://doi.org/10.1145/2857491.2857525.
[36] Horne, J.A. and Ostberg, O. 1976. A Self Assessment Questionnaire to Determine Morningness Eveningness in Human Circadian Rhythms. *International Journal of Chronobiology*. (1976).
[37] Huey, E.B. 1908. The Psychology and Pedagogy of Reading: With a Review of the History of Reading and Writing and of Methods, Texts, and Hygiene in Reading. (1908).
[38] Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E. and D'Mello, S.K. 2017. Gaze-based Detection of Mind Wandering during Lecture Viewing. *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)* (2017), 226–231.
[39] Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J.R. and D'Mello, S.K. 2019. Automated Gaze-Based Mind Wandering Detection during Computerized Learning in Classrooms. *User Modeling and User-Adapted Interaction*. 29, 4 (Sep. 2019), 821–867.https://doi.org/10.1007/s11257-019-09228-5.
[40] Hutt, S., Krasich, K., R. Brockmole, J. and K. D'Mello, S. 2021. Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. (2021).
[41] Jacob, R., Sellmach, S. and Stellmach, S. 2016. What You Look at Is What You Get: Gaze-Based User Interfaces. *Interactions*. 23, 5 (Aug. 2016), 62–65.https://doi.org/10.1145/2978577.
[42] Jung, C.G. 1915. Psychology of the Unconscious. (1915).
[43] Kar, A. and Corcoran, P. 2017. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*.
[44] Kruger, J.L., Hefer, E. and Matthew, G. 2013. Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. *ACM International Conference Proceeding Series* (2013).https://doi.org/10.1145/2509315.2509331.
[45] Kuniavsky, M. 2003. Observing the User Experience. *Observing the User Experience*. (2003).https://doi.org/10.1016/b978-1-55860-923-5.x5026-8.
[46] Lai, M.L., Tsai, M.J., Yang, F.Y., Hsu, C.Y., Liu, T.C., Lee, S.W.Y., Lee, M.H., Chiou, G.L., Liang, J.C. and Tsai, C.C. 2013. A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*.
[47] Lallé, S., Conati, C. and Carenini, G. 2016. Predicting confusion in information visualization from eye tracking and interaction data. *IJCAI International Joint Conference on Artificial Intelligence* (2016).
[48] Land, M. and Tatler, B. 2012. Looking and Acting: Vision and Eye Movements in Natural Behaviour. *Looking and Acting: Vision and Eye Movements in Natural Behaviour*. (2012).https://doi.org/10.1093/acprof:oso/9780198570943.001.0001.
[49] Lewis, J.R. 1995. Computer System Usability Questionnaire. *International Journal of Human-Computer Interaction*. (1995).https://doi.org/10.1037/t32698-000.
[50] Lin, Y.T., Lin, R.Y., Lin, Y.C. and Lee, G.C. 2013. Real-Time Eye-Gaze Estimation Using a Low-Resolution Webcam. *Multimedia Tools and Applications*. 65, 3 (2013), 543–568.https://doi.org/10.1007/s11042-012-1202-1.
[51] Liu, A., Xia, L., Duchowski, A.T., Bailey, R., Holmqvist, K. and Jain, E. 2019. Differential Privacy for Eye-Tracking Data. *Eye Tracking Research and Applications Symposium (ETRA)*. (2019).https://doi.org/10.1145/3314111.3319823.
[52] Al Madi, N., Peterson, C.S., Sharif, B. and Maletic, J.I. 2021. From Novice to Expert: Analysis of Token Level Effects in a Longitudinal Eye Tracking Study. *2021 IEEE/ACM 29th International Conference on Program Comprehension (ICPC)* (2021), 172–183.https://doi.org/10.1109/ICPC52881.2021.00025.
[53] Malhotra, N.K., Kim, S.S. and Agarwal, J. 2004. Internet users' information privacy concerns (IUIPC): The construct,

the scale, and a causal model. *Information Systems Research*.

[54] Mark, G., Iqbal, S., Czerwinski, M. and Johns, P. 2014. Capturing the mood: Facebook and face-to-face encounters in the workplace. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2014).https://doi.org/10.1145/2531602.2531673.

[55] Mark, G., Iqbal, S.T., Czerwinski, M., Johns, P. and Sano, A. 2016. Email duration, batching and self-interruption: Patterns of email use on productivity and stress. *Conference on Human Factors in Computing Systems - Proceedings* (2016).https://doi.org/10.1145/2858036.2858262.

[56] Martinez, G.J., Mattingly, S.M., Young, J., Faust, L., Dey, A.K., Campbell, A.T., De Choudhury, M., Mirjafari, S., Nepal, S.K., Robles-Granda, P., Saha, K. and Striegel, A.D. 2020. Improved Sleep Detection Through the Fusion of Phone Agent and Wearable Data Streams. (2020).https://doi.org/10.1109/percomworkshops48775.2020.9156211.

[57] Mattingly, S.M. et al. 2019. The Tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. *Conference on Human Factors in Computing Systems - Proceedings* (2019).https://doi.org/10.1145/3290607.3299041.

[58] Maurer, B., Krischkowsky, A. and Tscheligi, M. 2017. Exploring Gaze and Hand Gestures for Non-Verbal In-Game Communication. *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play* (2017), 315–322.https://doi.org/10.1145/3130859.3131296.

[59] Meng, L., Miao, C. and Leung, C. 2017. Towards Online and Personalized Daily Activity Recognition, Habit Modeling, and Anomaly Detection for the Solitary Elderly through Unobtrusive Sensing. *Multimedia Tools and Applications*. (2017).https://doi.org/10.1007/s11042-016-3267-8.

[60] Mills, C., Bixler, R., Wang, X. and D'Mello, S.K. 2016. Automatic gaze-based detection of mind wandering during film viewing. *The 9th International Conference on Educational Data Mining*. (Raleigh, North Carolina, 2016), 30–37.

[61] Mills, C., Gregg, J.M., Bixler, R. and D'Mello, S.K. 2021. Eye-Mind Reader: An Intelligent Reading Interface That Promotes Long-Term Comprehension by Detecting and Responding to Mind Wandering. *Hum. Comput. Interact.* 36, 4 (2021), 306–332.https://doi.org/10.1080/07370024.2020.1716762.

[62] Mundnich, K., Booth, B.M., l'Hommedieu, M., Feng, T., Girault, B., L'hommedieu, J., Wildman, M., Skaaden, S., Nadarajan, A., Villatte, J.L. and others 2020. TILES-2018, a Longitudinal Physiologic and Behavioral Data Set of Hospital Workers. *Scientific Data*. 7, 1 (2020), 1–26.

[63] Nasiopoulos, E., Risko, E.F., Foulsham, T. and Kingstone, A. 2015. Wearable Computing: Will It Make People Prosocial? *British Journal of Psychology*. 106, 2 (2015), 209–216.https://doi.org/10.1111/bjop.12080.

[64] Navarro, D. and Sundstedt, V. 2017. Simplifying game mechanics: gaze as an implicit interaction method. *SIGGRAPH Asia 2017 Technical Briefs* (2017), 4.https://doi.org/10.1145/3145749.3149446.

[65] Nyström, M., Andersson, R., Holmqvist, K. and Van De Weijer, J. 2013. The Influence of Calibration Method and Eye Physiology on Eyetracking Data Quality. *Behavior Research Methods*. 45, 1 (2013), 272–288.

[66] O'Brien, H.L., Cairns, P. and Hall, M. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human Computer Studies*. 112, July 2017 (2018), 28–39.https://doi.org/10.1016/j.ijhcs.2018.01.004.

[67] Özdemir, A.T. and Barshan, B. 2014. Detecting Falls with Wearable Sensors Using Machine Learning Techniques. *Sensors*. 14, 6 (2014), 10691–10708.

[68] Palumbo, R. V., Marraccini, M.E., Weyandt, L.L., Wilder-Smith, O., McGee, H.A., Liu, S. and Goodwin, M.S. 2017. Interpersonal Autonomic Physiology: A Systematic Review of the Literature. *Personality and Social Psychology Review*. 21, 2 (2017), 99–141.https://doi.org/10.1177/1088868316628405.

[69] Parak, J. and Korhonen, I. 2014. Evaluation of wearable consumer heart rate monitors based on photopletysmography. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014* (2014).https://doi.org/10.1109/EMBC.2014.6944419.

[70] Paul, C., Scheibe, K. and Nilakanta, S. 2020. Privacy Concerns Regarding Wearable IoT Devices: How it is Influenced by GDPR? *Proceedings of the 53rd Hawaii International Conference on System Sciences* (2020), 4388–4397.https://doi.org/10.24251/hicss.2020.536.

[71] Proudfoot, M., Menke, R.A.L., Sharma, R., Berna, C.M., Hicks, S.L., Kennard, C., Talbot, K. and Turner, M.R. 2016. Eye-Tracking in Amyotrophic Lateral Sclerosis: A Longitudinal Study of Saccadic and Cognitive Tasks. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*. 17, 1–2 (2016), 101–111.https://doi.org/10.3109/21678421.2015.1054292.

[72] Purta, R., Mattingly, S.M., Song, L., Lizardo, O., Hachen, D., Poellabauer, C. and Striegel, A. 2016. Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits. *International Symposium on Wearable Computers, Digest of Papers* (2016).https://doi.org/10.1145/2971763.2971767.

[73] Rahal, R.-M. and Fiedler, S. 2019. Understanding Cognitive and Affective Mechanisms in Social Psychology through Eye-Tracking. *Journal of Experimental Social Psychology*. 85, (2019), 103842.

[74] Ramadan, S., Abd-almageed, W. and Smith, C.E. 2002. Eye Tracking Using Active Deformable Models. December (2002).

[75] Raney, G.E., Campbell, S.J. and Bovee, J.C. 2014. Using Eye Movements to Evaluate the Cognitive Processes Involved in Text Comprehension. *Journal of Visualized Experiments*. (2014).https://doi.org/10.3791/50780.

[76] Rayner, K. 2009. Eye Movements and Attention in Reading, Scene Perception, and Visual Search. *Quarterly Journal of Experimental Psychology*. 62, 8 (2009).https://doi.org/10.1080/17470210902816461.

[77] Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*. 124, 3 (Nov. 1998), 372–422.https://doi.org/10.1037/0033-2909.124.3.372.

[78] Reinders, M.J.T. 2014. Eye Tracking by Template Matching Using an Automatic Codebook Generation Scheme. January 1997 (2014).
[79] Rigas, I., Komogortsev, O. and Shadmehr, R. 2016. Biometric Recognition via Eye Movements: Saccadic Vigor and Acceleration Cues. *ACM Transactions on Applied Perception*. 13, 2 (2016).https://doi.org/10.1145/2842614.
[80] Risko, E.F. and Kingstone, A. 2017. Everyday attention. *Canadian Journal of Experimental Psychology*.
[81] Ross, R.G., Harris, J.G., Olincy, A. and Radant, A. 2000. Eye Movement Task Measures Inhibition and Spatial Working Memory in Adults with Schizophrenia, ADHD, and a Normal Comparison Group. *Psychiatry Research*. 95, 1 (2000), 35–42.
[82] Russell, J.A. 2003. Core Affect and the Psychological Construction of Emotion. *Psychological Review*. 110, 1 (2003), 145.
[83] Saha, K., Chan, L., De Barbaro, K., Abowd, G.D. and De Choudhury, M. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. (2017).https://doi.org/10.1145/3130960.
[84] Shany, T., Redmond, S.J., Narayanan, M.R. and Lovell, N.H. 2011. Sensors-Based Wearable Systems for Monitoring of Human Movement and Falls. *IEEE Sensors Journal*. 12, 3 (2011), 658–670.
[85] Smyth, C. 2000. The Pittsburgh Sleep Quality Index (PSQI). *Director (Cincinnati, Ohio)*. (2000).https://doi.org/10.1067/min.2000.107649.
[86] SRRESEARCH 2016. *The EyeLink ® 1000 Plus Eye Tracker One Camera , Many Different Eye Tracking Solutions*.
[87] Steil, J., Hagestedt, I., Huang, M.X. and Bulling, A. 2019. Privacy-Aware Eye Tracking Using Differential Privacy. *Eye Tracking Research and Applications Symposium (ETRA)*. (2019).https://doi.org/10.1145/3314111.3319915.
[88] Steindorf, L. and Rummel, J. 2020. Do Your Eyes Give You Away? A Validation Study of Eye-Movement Measures Used as Indicators for Mindless Reading. *Behavior Research Methods*. (2020).https://doi.org/10.3758/s13428-019-01214-4.
[89] Tien, T., Pucher, P.H., Sodergren, M.H., Sriskandarajah, K., Yang, G.-Z. and Darzi, A. 2014. Eye Tracking for Skills Assessment and Training: A Systematic Review. *Journal of Surgical Research*. 191, 1 (2014), 169–178.https://doi.org/https://doi.org/10.1016/j.jss.2014.04.032.
[90] Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K. and Navalpakkam, V. 2020. Accelerating Eye Movement Research via Accurate and Affordable Smartphone Eye Tracking. *Nature Communications*. 11, 1 (2020), 4553.https://doi.org/10.1038/s41467-020-18360-5.
[91] Vosskuhler, A., Nordmeier, V., Kuchinke, L. and Jacobs, A.M. 2008. OGAMA (Open Gaze and Mouse Analyzer): Open-Source Software Designed to Analyze Eye and Mouse Movements in Slideshow Study Designs. *Behavior Research Methods*. 40, 4 (Nov. 2008), 1150–1162.https://doi.org/10.3758/BRM.40.4.1150.
[92] Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D. and Campbell, A.T. 2014. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014).https://doi.org/10.1145/2632048.2632054.
[93] Wedel, M. and Pieters, R. 2008. A Review of Eye-Tracking Research in Marketing. *Review of Marketing Research*. (2008).https://doi.org/10.1108/S1548-6435(2008)0000004009.
[94] Weibel, N., Fouse, A., Emmenegger, C., Kimmich, S. and Hutchins, E. 2012. Let's look at the cockpit: exploring mobile eye-tracking for observational research on the flight deck. *Proceedings of the Symposium on Eye Tracking Research and Applications* (2012), 107–114.https://doi.org/10.1145/2168556.2168573.
[95] Woll, S. 2001. Everyday Thinking: Memory, Reasoning, and Judgment in the Real World. (2001).
[96] Wu, Y.L., Yeh, C.T., Hung, W.C. and Tang, C.Y. 2014. Gaze Direction Estimation Using Support Vector Machine with Active Appearance Model. *Multimedia Tools and Applications*. 70, 3 (2014), 2037–2062.https://doi.org/10.1007/s11042-012-1220-z.
[97] Yonetani, R., Kawashima, H. and Matsuyama, T. 2012. Multi-mode Saliency Dynamics Model for Analyzing Gaze and Attention. *Proceedings of the Symposium on Eye Tracking Research and Applications* (New York, NY, USA, 2012), 115–122.https://doi.org/10.1145/2168556.2168574.
[98] Zhang, X., Sugano, Y., Fritz, M. and Bulling, A. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 41, 1 (2019), 162–175.https://doi.org/10.1109/TPAMI.2017.2778103.
[99] Zhang, Y., Chong, M.K., Müller, J., Bulling, A. and Gellersen, H. 2015. Eye Tracking for Public Displays in the Wild. *Personal and Ubiquitous Computing*. 19, 5 (2015), 967–981.https://doi.org/10.1007/s00779-015-0866-8.

Received November 2021; revised January 2022; accepted April 2022.

[1] https://www.sr-research.com/

[2] The Tobii 4C requires an additional research license in order to be used for recording purposes, which we purchased for each of the units used in this study

[3] https://developer.tobii.com/pc-gaming/unity-sdk/api-overview/

[4] Code for both data sampling and transfer is available on github, https://github.com/emotive-computing/ETRA_2022