Behavior Research Methods
https://doi.org/10.3758/s13428-023-02143-z

# Investigating the suitability of online eye tracking for psychological research: Evidence from comparisons with in-person data using emotion–attention interaction tasks

**Paul C. Bogdan[1,2] · Sanda Dolcos[1,2] · Simona Buetti[2] · Alejandro Lleras[1,2] · Florin Dolcos[1,2,3]**

Accepted: 9 May 2023


**Abstract**
The future is bound to bring rapid methodological changes to psychological research. One such promising candidate is the use of webcam-based eye tracking. Earlier research investigating the quality of online eye-tracking data has found increased spatial and temporal error compared to infrared recordings. Our studies expand on this work by investigating how this spatial error impacts researchers' abilities to study psychological phenomena. We carried out two studies involving emotion–attention interaction tasks, using four participant samples. In each study, one sample involved typical in-person collection of infrared eye-tracking data, and the other involved online collection of webcam-based data. We had two main findings: First, we found that the online data replicated seven of eight in-person results, although the effect sizes were just 52% [42%, 62%] the size of those seen in-person. Second, explaining the lack of replication in one result, we show how online eye tracking is biased toward recording more gaze points near the center of participants' screen, which can interfere with comparisons if left unchecked. Overall, our results suggest that well-powered online eye-tracking research is highly feasible, although researchers must exercise caution, collecting more participants and potentially adjusting their stimulus designs or analytic procedures.

**Keywords** Gaze · Affect · Perception · Emotion-cognition interaction · Emotion regulation

## Introduction

Eye tracking is a valuable tool for psychological research, and its utility for a wide variety of fields (e.g., decision-making) is becoming ever more apparent (Amasino et al., 2019; Kragel & Voss, 2022; Strohmaier et al., 2020; Voss et al., 2017). However, infrared eye-tracking systems remain logistically prohibitive, particularly for researchers who acknowledge the benefits of eye tracking but do not use it as one of their primary methodologies. Fortunately, over the past few years, several groups have released freely available software packages that allow eye-tracking data to be recorded via webcam (e.g., *WebGazer.js*; Papoutsaki et al., 2016), and a handful of previous studies have explored the efficacy of webcam eye tracking (Schneegans et al., 2021; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021; Slim & Harsuiker, 2022; Vos et al., 2022; Degen et al., 2021). These studies generally agree that webcam eye tracking replicates some patterns found in person, but it limits data quality. Building on this work, we pursued two complementary lines of questions: (1) How much worse is webcam eye tracking in terms of detecting the effects of experimental manipulations, and is it still feasible and useful to collect the sample size needed for a well-powered webcam eye-tracking study? (2) Why may some results not replicate online, and can online noise introduce biases into specific analyses? These directions were specifically chosen to best assess the suitability of this technology for psychological research.

✉ Paul C. Bogdan
pbogda2@illinois.edu

✉ Florin Dolcos
fdolcos@illinois.edu

1 Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, IL, USA

2 Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

3 Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, IL, USA

## Data quality and spatial error

Earlier studies on the quality of online eye tracking have focused on its spatial and temporal error. *Webgazer.js*'s spatial error is generally 3–4.5° of visual angle for online participants (Papoutsaki et al., 2016; Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022). High spatial error is often linked to low-quality webcams, which can vary widely in levels of detail (e.g., 0.3–2.1 megapixels). In contrast, infrared systems achieve spatial accuracy on the order of 0.5° visual angle (Ehinger et al., 2019). For typical viewing distances, 3–4.5° corresponds to 15–20% of participants' screens. Such percentages would severely limit designs where stimuli are close together. However, if participants are asked to sit closer to their screens, 3–4.5° would amount to a much smaller portion of it (e.g., 5%). Sitting closer to one's screen carries drawbacks, such as causing participants to turn their heads when moving their eyes. Nonetheless, we opted for this strategy, as our studies benefit from higher spatial resolution. Webcam-based eye tracking is also often linked to a delay compared to infrared data, being roughly 400 ms slower in detecting saccades (Degen et al., 2021; Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022). Delays may create particular challenges for designs requiring rapid changes in attention. Our studies are largely not impacted by delays.

Although spatial error has been quantified in terms of the distances between recorded vs. actual gaze, it remains unclear how this impacts the aggregate-level quantities that researchers are primarily concerned with. For example, how does this noise impact calculations on the time participants spent gazing within an interest area? Work by Schneegans et al. (2021) begins to shed light topic. In their decision-making task, participants chose between two options while eye tracking was recorded. In-person participants were recorded as spending 75% of the decision time gazing toward the option they selected (25% toward non-selected option), whereas online participants were recorded as spending just 63% toward the option they selected (37% toward non-selected option). Thus, the interest area time seems to contract toward 50%, which is intuitive, since infinite error would lead to exactly 50% per interest area if both areas were equally sized.

However, contraction may not fully represent the impact of spatial error on aggregate patterns. Given the heatmaps reported by previous studies, online eye tracking seems to elicit a "centering bias" whereby more gaze points are recorded near the center of participants' screens. For example, see the fourth figure of Semmelmann and Weigelt (2018), the fourth figure of Slim and Hartsuiker (2022), and the fifth figure of Vos et al. (2022). No previous online eye-tracking report has investigated this centering bias, although it may impact the analysis of psychological effects on gaze, much like contraction pattern.

## Replication of in-person results

Several online eye-tracking studies have attempted to replicate gaze results from in-person experiments (Degen et al., 2021; Schneegans et al., 2021; Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022; Yang & Krajbich, 2021). Each study showed how some or all of the in-person results could be replicated online, although the online results were dulled and required larger sample sizes. The more recent study by Slim and Hartsuiker (2022) attempted to quantify the extent to which dulling occurred by examining the change in effect sizes and estimating how many additional participants would be required for online research.

Examining effect sizes is a means of assessing data quality that departs from other previous papers, which have instead focused on spatial or temporal resolution. Measurements of spatial error are valuable but difficult to interpret and apply for the ultimate questions many readers will have: Will online eye tracking be good enough for testing their hypotheses? How many additional participants do researchers need to collect to overcome the drop in data quality? Examining effect sizes strikes the core of these questions. Slim and Hartsuiker (2022) did this but only replicated a single effect and so their estimated change in effect sizes comes with a wide confidence interval and limited generalizability. Precision is invaluable for efficiently assessing necessary sample sizes. Accordingly, the present research attempted to replicate multiple results (more than most of these earlier replication studies combined) covering a range of effect sizes, including both large ($d > 0.8$) and small-to-moderate ($0.4 < d < 0.5$). Furthermore, we conducted a meta-analysis, aggregating our findings on the effect sizes to be expected to generate a more exact estimate.

## Present study

In sum, the present study investigated the quality of online eye-tracking data with a focus on aggregate patterns and the ability to detect gaze-related effects in studies, as this direction best assesses the suitability of this technology for psychological research. We investigated these questions using data from two studies on emotion–attention interactions. For each study, one data set was collected in person, while eye tracking was recorded via a typical infrared system, and the other data set was collected online, while eye tracking was recorded via participants' webcams. The two studies differed slightly in their tasks. Study 1 involved presenting participants with a series of negative and neutral images, which they freely examined. This study built upon earlier research on the salience of emotional information (Carretié, 2014) and the wide-reaching links between emotion and attention (Dolcos, Katsumi, et al., 2020b). Study 2 involved presenting participants with the same series of images, but with instructions on how to attend the images. This study concerned how individuals can use top-down attentional control to regulate their emotions (Gross & John, 2003). Notably, Study 2 partially relied on recently published in-person data (Dolcos et al., 2022), and the analyses followed identical procedures.

We tested whether each effect seen in-person was replicated online and measured the extent to which effect sizes dipped when transitioning to online collection. To achieve a precise estimate of the effect size drop, the results below also include a meta-analysis, combining each effect size drop seen into a single precise measurement. To further understand in-person vs. online differences, simulations were also conducted. The simulations modeled the recorded online data as a function of the in-person data plus added Gaussian noise. We expected that this would recreate the aggregate online patterns, including the centering bias postulated above. Through the simulations, we also modeled how the centering bias may impact analyses.

## Study 1

Emotional stimuli are known to capture attention. For example, when participants are shown an image containing both negative and neutral elements, they spend more time looking at the negative parts of the image (Carretié, 2014; Öhman et al., 2001). Our first study was based on this principle. While eye tracking was recorded, participants viewed a series of composite images containing a negative or neutral foreground overlaid upon a neutral background, with no specific instructions on how to scan the images (free viewing). Following the presentation of each image, participants reported the intensity of their emotional reactions. We expected to (1) find that participants spent more time gazing within the foregrounds for negative images relative to neutral images. Furthermore, we expected to (2) identify trial-by-trial links between foreground time and participants' emotional ratings of the images. We predicted that these results would emerge in both in-person and online data.

### Methods

#### Participants

We recruited 139 participants from the local university (84 female; 53 male; $M_{age} = 19.7$; $SD_{age} = 1.30$) for this study. Thirty-four participants completed the in-person version of the task (none excluded), and 105 completed the online version. The in-person data were collected pre-pandemic when almost all subject pool studies were in-person, and the online data were collected during the pandemic when all subject pool studies were online. From the online sample, 18 participants were excluded, as they either responded to fewer than 75% of the emotional rating trials, showed extreme outlier responses in the emotional rating data for the neutral trials (Z > 5), or fewer than 75% of their trials yielded usable eye-tracking data (no gaze recorded for any time point). Power analyses (α = .05, power = 80%) based on preliminary in-person data ($d_z = 0.72$) revealed that a minimum of 18 participants would be needed to identify emotion vs. neutral effects on in-person gaze, and 63 participants would be needed to identify online effects, assuming a 50% drop in effect size. However, additional participants were recruited based on memory-related hypotheses, although these were not the current focus. Sensitivity analyses revealed that the 34 in-person participants were sufficient to detect effects at least $d_z = 0.50$, and the 87 online participants were sufficient to detect effects at least $d_z = 0.31$. All participants provided informed consent under a protocol approved by the University of Illinois Institutional Review Board and received course credit in exchange for participation.

#### Task design

Eye movements were recorded while participants viewed a series of 90 composite images (60 negative and 30 neutral; Fig. 1). Each composite image was created by overlaying a negative or neutral foreground component upon a visually complex neutral background, such that the image was approximately 50% foreground and 50% background. The foreground components were extracted from images part of the International Affective Picture System (IAPS; Lang et al., 2008), the Geneva Affective Picture Database (GAPED; Dan-Glauser & Scherer, 2011), the Military Affective Picture System (MAPS; Goodman et al., 2016), the Nencki Affective Picture System (NAPS; Marchewka et al., 2014), and the Emotional Picture Set (EmoPicS; Wessa et al., 2010). These pictures sets and freely available online sources were used for the BG components. Negative and neutral composite images were matched for foreground location (i.e., top, bottom, left, right, middle), complexity, brightness, contrast, human presence, and animacy (e.g., animals vs. objects; all *ps* > .05). A validation study (*N* = 19) using nine-point Likert scales confirmed that the emotional images were negatively valenced ($M_{Valence} = 2.46$, $SD_{Valence} = 0.79$) and arousing ($M_{Arousal} = 4.95$, $SD_{Arousal} = 1.05$), while neutral images were appropriately neutral ($M_{Valence} = 4.79$, $SD_{Valence} = 0.48$) and non-arousing ($M_{Arousal} = 2.17$, $SD_{Arousal} = 0.46$).

Each composite image was presented for 4 s. Images were followed by a Likert scale asking participants to rate their emotional reaction to each image (1 = "*Not at all negative*"; 5 = "*Very negative*"). Testing for possible group differences reveals that online participants reported numerically lower emotional ratings in the Emotion condition (M = 3.31) than


**Fig. 1** Task diagram for Study 1. Participants were instructed to look at the upcoming image and rate their emotional response to the image from 1 ("not at all negative") to 5 ("very negative). The images all had a composite structure with roughly 50% dedicated to foreground content (which could be either negative or neutral) and 50% to neutral background content

did the in-person participants (M = 3.50; $t$[120] = – 1.82, $p$ = .07). This likely reflects lower engagement for the online study, but both groups still showed far higher ratings than in the Neutral conditions ($M_{online}$ = 1.25, $M_{in\text{-}person}$ = 1.26). Thus, we expected the differences in engagement to only have a small impact, which we will discuss later. Both studies also collected self-reported data using questionnaires (see Supplemental Materials 1), but those data were not the current focus. A subset of participants also completed a delayed memory task, but this was also beyond the current focus.

## Eye-tracking procedures

For the in-person task, eye positions were recorded from each participant's right eye using the infrared EyeLink 1000 system (SR Research, ON, Canada). For the online task, eye positions were recorded using participants' webcams via the *WebGazer.js* package (https://webgazer.cs.brown.edu/; Papoutsaki et al., 2016). Complete details on the in-person and online eye-tracking procedures and calibration are provided in Supplemental Materials 2. *WebGazer.js* generates a coordinate ($x$, $y$) at each time point. Unlike most infrared eye-tracking software, *WebGazer.js* does not automatically detect blinks for exclusion. Rather, *WebGazer.js* roughly records the coordinate where participants were gazing prior to the blink.

The presented composite images were large (6" height x 8" width on a 15.6" laptop screen) and spanned approximately 50% of participants' screens. Additionally, participants were instructed to sit closer to their laptop screens (12–15"), as pilot testing showed that this distance improved data quality. Hence, the pictures captured approximately 28° and 36° of participants' vertical and horizontal visual fields, respectively. These large sizes were expected to enhance the quality of the eye tracking, as errors during recording would have relatively smaller impacts.

Analysis of the calibration data showed that the average degree of spatial error (root mean square) was 4.8% of the screen's width in the $x$-dimension, which corresponds to 0.65" and 2.8° visual angle on a standard laptop screen (15.6" diagonal and 19:6 aspect ratio). The error was 5.9% of the screen's height in the $y$-dimension, which corresponds to 0.45" and 1.9° visual angle. The total Euclidean distance is 0.79" and 3.3°, which is in line with the error reported by earlier webcam-based eye-tracking studies (Papoutsaki et al., 2016; Semmelmann & Weigelt, 2018). These distances are consistent with participants' subjective assessment of error. Before the task, participants were shown their predicted gaze location in real time with a moving dot and asked to judge its precision. Twelve percent of participants reported that the dot distance was less than 0.5", 33% reported that it was 0.5–1", 42% reported 1–1.5", 8% reported 1.5–2", and 4% reported a distance greater than 2". The average sampling rate was 33.1 Hz. The sampling rate varied highly between participants (SD = 15.8 Hz) but was stable within-participant (between-trial SD = 1.3 Hz).

For the in-person group, gaze points recorded as outside the image made up 2% of gaze time, with 0.7% of gaze time recorded as off-screen entirely or as dead periods. For the online group, gaze points outside the image made up 13% of gaze time, with 5% recorded as off-screen (*WebGazer.js* does report dead periods). Because such gaze points outside the image likely reflect a brief malfunctioning of the eye tracker, they were ignored from analyses, and thus the total proportion spent in the foreground or background components summed to 100%. The online data is too noisy to effectively assess fixations. As in previous online eye-tracking studies (Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021), the gaze data were analyzed in terms of the time spent within a given interest area.

### Analytic procedures

The proportions of gaze points recorded in the foreground were submitted to paired *t* tests and multilevel linear regressions. The *t* tests examined the effects of image type (emotion vs. neutral) on gaze. The multilevel linear regressions measured trial-by-trial links between gaze and emotional rating. The full regression equations are provided in Supplemental Materials 3. All of the multilevel regressions followed recommended practices (Meteyard & Davies, 2020), including the use of a maximal random effects structure (Barr et al., 2013), restricted maximum likelihood (Luke, 2017), and Satterthwaite's method of estimating degrees of freedom (Luke, 2017). These procedures are known to minimize the type I error rate (Barr et al., 2013; Luke, 2017; Meteyard & Davies, 2020). The multilevel regressions were fit using *R* (R Core Team, 2013) and the *lme4* package (Bates et al., 2014). Fixed effect significance was calculated using the *lmerTest* package (Kuznetsova et al., 2017). For the *t* tests, effect sizes were calculated as Cohen's $d_z$ (*t* value divided by the square root of the sample size); $d_z$ was well suited for comparing the in-person and online data sets, as it is typically used for power analyses and estimating the sample sizes needed to detect a given result (Lakens, 2013). Measures of "$d_z$" were also calculated for the multilevel regression effect sizes, based on the fixed effect *t* values. Although this is not standard procedure for multilevel model effect sizes (Lorah, 2018), using "$d_z$" was effective because it allows seamless comparisons between the in-person and online effects in terms of the statistical power researchers can expect when transitioning to online collection.

### Simulation and centering bias

We expected the online data's heatmap, averaged across both conditions, to show a centering bias. We simulated whether this centering bias may be rooted in the heightened spatial error that necessarily accompanies online eye tracking (see Fig. 2 for a visual explanation). To confirm this link, our simulation added Gaussian noise to the in-person gaze time series on a point-by-point basis,

$$x_i' = x_i + e,$$

$$e \sim N(0, 0.1),$$

where $x_i$ is an original in-person gaze point, $e$ is noise sampled from a normal distribution, and $x_i'$ is the resulting measurement, which simulates the noisy online data. The amount of Gaussian noise (σ = 0.10) was selected as the smallest amount necessary to simulate the attentional-aversion flip (see below). The simulation was carried out ten times, then averaged.

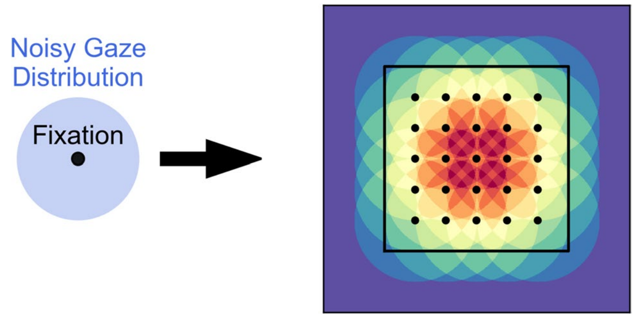


**Fig. 2** Spatial noise can cause a centering bias. The element on the *left* represents how a fixation would be recorded by a webcam eye tracker, and the *right panel* shows how this noise leads to more gaze points recorded near the center of the image. For simplicity, the noise is represented as a *circle*, but the point holds if instead it is represented by any shape

## Results

### Replication

Consistent with our predictions, clear attention-capturing effects of emotion emerged in both the in-person and online eye-tracking studies. Specifically, participants spent significantly more time in the foreground areas of negative than neutral images, both in the in-person ($t$[33] = 6.65, $p$ < .001, $d_z$ = 1.14; Fig. 3A) and online data ($t$[86] = 5.72, $p$ < .001, $d_z$ = 0.61; Fig. 3B).

The trial-by-trial analyses using multilevel regressions also yielded robust replications. Analogous to the *t* test above, a regression predicting emotional ratings as a function of foreground gaze, showed strong effects for both the in-person ($\beta$ [standardized] = .20, $p$ < .001, $d_z$ = 0.99; Fig. 3C) and online data ($\beta$ = .07, $p$ < .001, $d_z$ = 0.524; Fig. 3D). Furthermore, a version of this regression, which controlled for whether the image was emotional vs. neutral, also showed significant effects of foreground gaze on ratings for both the in-person ($\beta$ = .042 $p$ = .01, $d_z$ = 0.45) and online data ($\beta$ = .02, $p$ = .03, $d_z$ = 0.23). This final regression demonstrates that gaze behavior dissociates between emotional images that elicit intense vs. mild reactions. This effect is a subtler difference than the initial emotional vs. neutral comparisons, and it shows the effectiveness of online eye tracking in identifying even relatively fine-grain patterns. Overall, these results suggest that the switch to online eye tracking led to decreased,

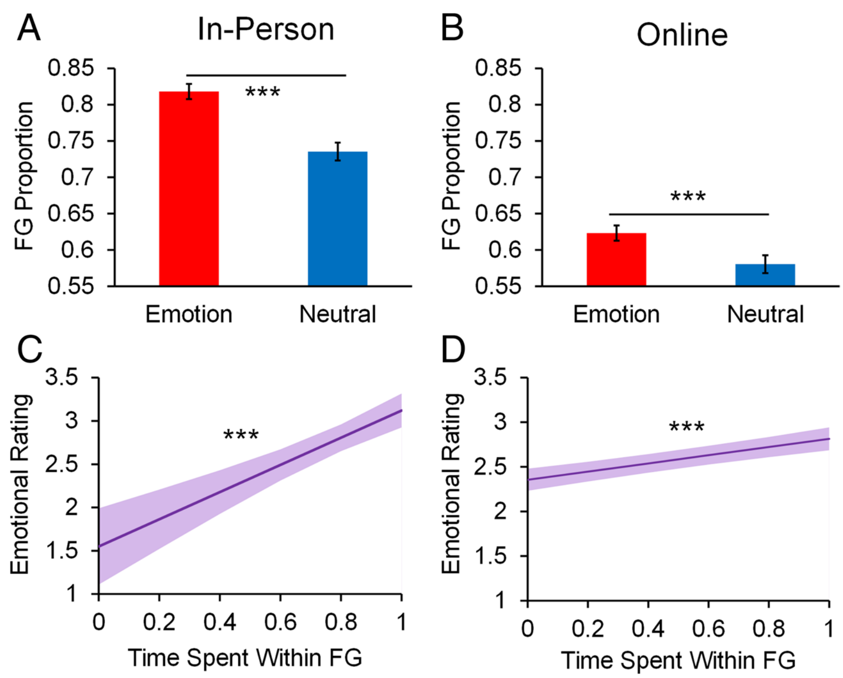


**Fig. 3** In-person vs. online comparison of the Study 1 results. **A** Representing the emotional vs. neutral paired *t* test on the in-person data. **B** Representing the same *t* test but on the online data. **C** Representing the multilevel regression linking gaze and emotional rating in the in-person data. **D** Representing the same regression but for the online data. *Error bars* represent 1 standard deviation above/below the mean. The *shaded region* represents a 95% confidence interval. *FG* foreground. ***, $p < .001$

albeit manageable effect sizes. For robustness, this point was further examined using data from another study.

### Simulation and centering bias

To further investigate the aggregate-level effects of transitioning to online eye tracking, we examined the heatmaps, averaged across both conditions, for each group. Whereas the in-person data have a largely uniform distribution (Fig. 4A), the online data shows a peak in the center (Fig. 4B). Simulating the transition from in-person to online eye tracking as the addition of Gaussian noise creates the same center-biased pattern (Fig. 4C). Thus, spatial noise, which may be symmetric in its impact on individual gaze points, can create emergent patterns when aggregated, like this centering bias.

## Study 2

Our second study aimed to replicate previously published results on emotion–attention interactions (Dolcos et al., 2022), which examined "focused attention" as an emotion-regulation strategy. Participants viewed the same series of negative and neutral images, as in Study 1, but now, before each image, participants were cued to focus on either its foreground or background areas. After each image, participants reported their emotional rating. The published eye-tracking data (Dolcos, Katsumi, et al., 2020a) showed that (1) participants could effectively control their attention based on the cues, focusing on the foreground or background content as instructed, but (2) some minor emotional attention-capture or attention-aversion effects were still present. Although

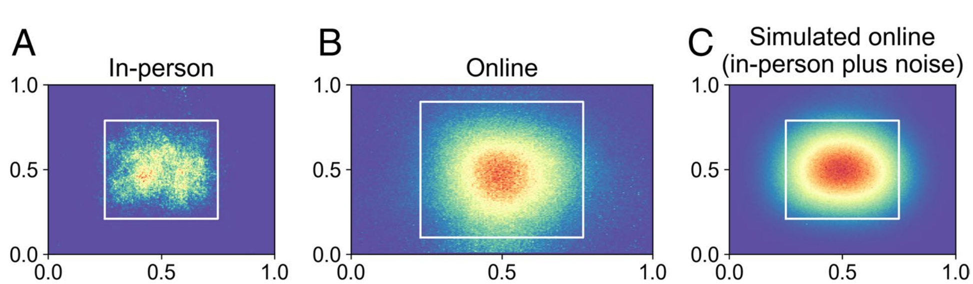


**Fig. 4** Heatmaps averaged across all images. The three heatmaps correspond to the in-person (**A**), online (**B**), and simulated (**C**) data. The Cartesian plane (0-1) corresponds to the participant's full screen, where the aspect ratio differed between in-person (4:3) and online (typically 16:9)

these second results were subtle and mostly peripheral to the main focus, they showed that when participants were instructed to focus on the background content of images, slight attention-capture effects still occurred (i.e., the emotional foreground captured their gaze). Also, when instructed to focus on the foreground, subtle attentional-aversion reactions also occurred, where participants slightly shifted their focus away from the emotional foregrounds. Returning to the main direction of this published research, trial-by-trial analyses showed that (3) time spent gazing within the background (i.e., the extent that participants carried out focused attention effectively) predicted the intensity of their emotional responses. This final pattern even emerged when statistically controlling for the attentional cue. For example, when participants were instructed to focus on the background, spending 90% of the time gazing within the background led to milder emotional reactions than spending 80%. These findings speak to the importance of gaze patterns during emotion processing and suggest that attentional control is a suitable target for emotion regulation (Dolcos et al., 2020a; Dolcos et al., 2020c; Gross & John, 2003; Strauss et al., 2016). Hence, we expected that similar patterns would also emerge in the online data.

Expanding on the Study 1 simulation results, we also tested how the centering bias elicited by spatial noise can impact gaze analyses and replication. Specifically, we conducted our noisy simulation and submitted the data to one of the analyses we sought to replicate – namely, the attentional-aversion effect. We examined how the results changed following the addition of noise and how this change interacts with interest area positions.

## Methods

### Participants

We recruited 264 participants from the local university (159 female; 99 male; $M_{age} = 20.1$, $SD_{age} = 1.64$) for this study. Forty-five completed the in-person version of the task, and 219 completed the online version. Three in-person participants were excluded due to outlier responses or technical issues (as was done by Dolcos et al., 2022). Thirty-nine online participants were excluded following the criteria of Study 1. The in-person sample size was justified by the power analysis described by Dolcos et al. (2022). For the online version, effects from an independent data set were used to carry out a power analysis for the main result of interest (emotional ratings regressed on gaze while controlling attention), which revealed that 130 participants would be sufficient. However, like in Study 1, a larger data set was collected due to hypotheses regarding memory-related effects, which are beyond the current focus. Sensitivity analyses revealed that the 41 in-person participants were sufficient to detect effects at least $d_z = 0.45$, and the 180 online participants were sufficient to detect effects at least $d_z = 0.21$. All participants provided informed consent under a protocol approved by the University of Illinois Institutional Review Board and received course credit in exchange for participation.

### Task protocols

Study 2 modified the design of Study 1 to provide an attentional cue before each image, which instructed participants to either focus on foreground or background components of composite images (task diagram shown in Supplemental Fig. S1). Half of the images were preceded by a foreground focus cue (30 negative, 15 neutral) and the other half by a background focus cue (30 negative, 15 neutral). Neither the mean nor the variance of the emotional ratings significantly differed between the in-person vs. online data sets, for any condition ($p$s > .12). As in Study 1, examination of the emotional rating data suggests slightly lower engagement for the online group. When instructed to focus on the emotional foreground, online participants showed numerically lower emotion ratings (M = 3.51) than in-person participants (M = 3.74, $t[221] = 1.48$, $p = .14$, $d = 0.10$). When instructed to focus on the background, online participants showed numerically higher ratings (M = 2.56) than in-person participants (M = 2.41, $t[221] = 1.03$, $p = .30$, $d = 0.07$). Although neither test reached significance, these numerical patterns may suggest that online participants exerted less top-down attentional control.

Eye-tracking procedures were the same as for Study 1 (see Supplemental Materials 2). The analyses mirrored those from Dolcos et al. (2022), involving paired $t$ tests and multilevel regressions. The $t$ tests examined the effects of the attentional cues (foreground focus vs. background focus) or image type (emotion vs. neutral) on gaze. The multilevel regressions examined the effect of gaze on emotional ratings. Multilevel modeling followed the recommended practices described in Study 1 (Meteyard & Davies, 2020), and the full regression equations are provided in Supplemental Materials 3.

### Aggregating the effect sizes as percentages

After completing the analyses, the results were aggregated to generate an overall estimate of the effect size change between in-person and online collection. This was calculated as follows: First, for each online result, a distribution of $d_z$ was generated via bootstrapping – i.e., constructing new data sets via random sampling with replacement, then submitting the constructed data sets to identical analyses (10,000 bootstrap simulations). Second, for each analysis that was replicated, the $d_z$ distribution was transformed into a distribution of percentages via dividing by the in-person $d_z$. For example, if an analysis yielded in-person $d_z = 0.8$ and online $d_z = 0.5$, this suggests that online effect sizes tend to be 62.5% of those found in person. This calculation was performed for the entire distribution – e.g., if the 95% confidence interval for $d_z$ was [0.3, 0.7], this would correspond to a percentage confidence interval of [37.5%, 87.5%]. Third, a meta-analysis was done on the distributions of percentages by taking the product (geometric mean) of the percentage distributions. The figure showing the aggregation below illustrates this procedure. One effect size distribution was an outlier in terms of having an extremely small standard deviation. This small standard deviation was due to the in-person effect size being extremely large. Given the small standard deviation, it was omitted as it would otherwise dominate the aggregation, contributing more than all other effects combined.

### Simulation and centering bias

The simulation procedures for Study 1 were applied again but now to test how the spatial error impacts replication. The simulation was done for the attentional-aversion result, which compared foreground gaze time between the emotional foreground-focus and neutral foreground-focus conditions. Noise was added to the in-person data for the emotional and neutral foreground-focus conditions, and the resulting noisy data were submitted to the same comparison. This was expected to reproduce the *t* test results seen online.

Furthermore, to test how this analysis is influenced by the centering phenomenon that emerges from low spatial resolution, we examined the extent our stimuli were susceptible to the bias. Our stimuli – composite images – were photorealistic combinations of foreground and background components and, accordingly, the interest areas were intricate. Inevitably, each image foreground was unique in shape and extent of overlap with the center, meaning each trial's data were differently susceptible to the centering bias. Quantifying this point, we defined a Gaussian distribution at the screen center, and we measured the extent each image's foreground overlapped with that distribution (overlap weighed by the distribution's density at a given point). Then, to test how differences in center overlap may impact analyses, we split our original pool of 60 emotional images into two 30-image subsets (Fig. 5). The first subset was created by matching each neutral image to the emotional ones that were the most similar in terms of overlap, hence creating a "matched" set of 30 emotional images (M = 0.597), which mirrored the centrality of the 30 neutral images (M = 0.592). The second subset was the remaining 30 emotional images, which were more central on average (M = 0.742) than the neutral ones ("unmatched"). The simulation was then conducted twice, testing for an attentional-aversion effect between the neutral images vs. matched emotional images and the neutral vs. unmatched emotional images.

## Results

### Replication

As expected, the eye-tracking data reliably dissociated the foreground focus vs. background focus conditions. Comparing the two focus conditions after averaging across the emotional and neutral trials revealed robust findings with large effect sizes in both the in-person ($t$[41] = 42.9, $p < .001$, $d_z = 6.62$; Fig. 6A vs. 6C) and online data ($t$[179] = 12.3, $p < .001$, $d_z = 0.92$; Fig. 6B vs. 6D). However, in a relative sense, the online effect size was much lower than that of the in-person effect size ($d_z = 0.92$ vs. $d_z = 6.62$).

The two subtler branches of analysis (secondary findings by Dolcos et al., 2022) yielded partial replications. These analyses investigated the effect of emotion on each focus condition. Examining the effect of emotion on the background focus trials revealed attention-capture effects in person ($t$[41] = 4.57, $p < .001$, $d_z = 0.71$; Fig. 6A), which were replicated online ($t$[179] = 4.48, $p < .001$, $d_z = 0.35$; Fig. 6B). In other words, when asked to focus on the background, participants

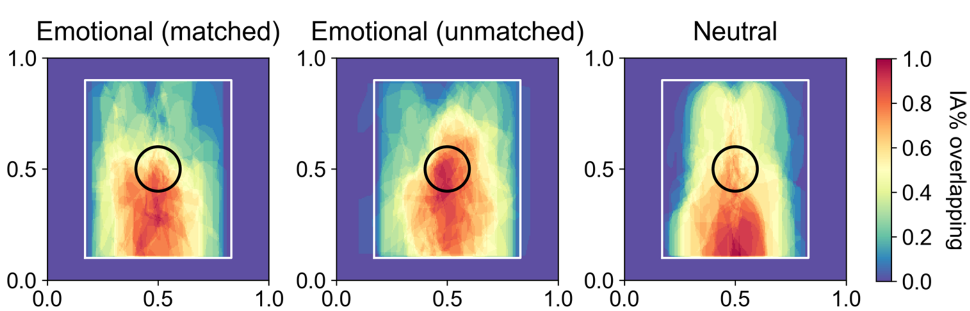


**Fig. 5** Interest areas for the emotional and neutral conditions. 1.0 (*hot*) indicates that every interest area (IA) overlaps a given screen pixel, whereas 0.0 (*cold*) indicates that no interest area overlaps it. *Black circles* were added as visual references to help compare the different stimulus categories' overlap with the center

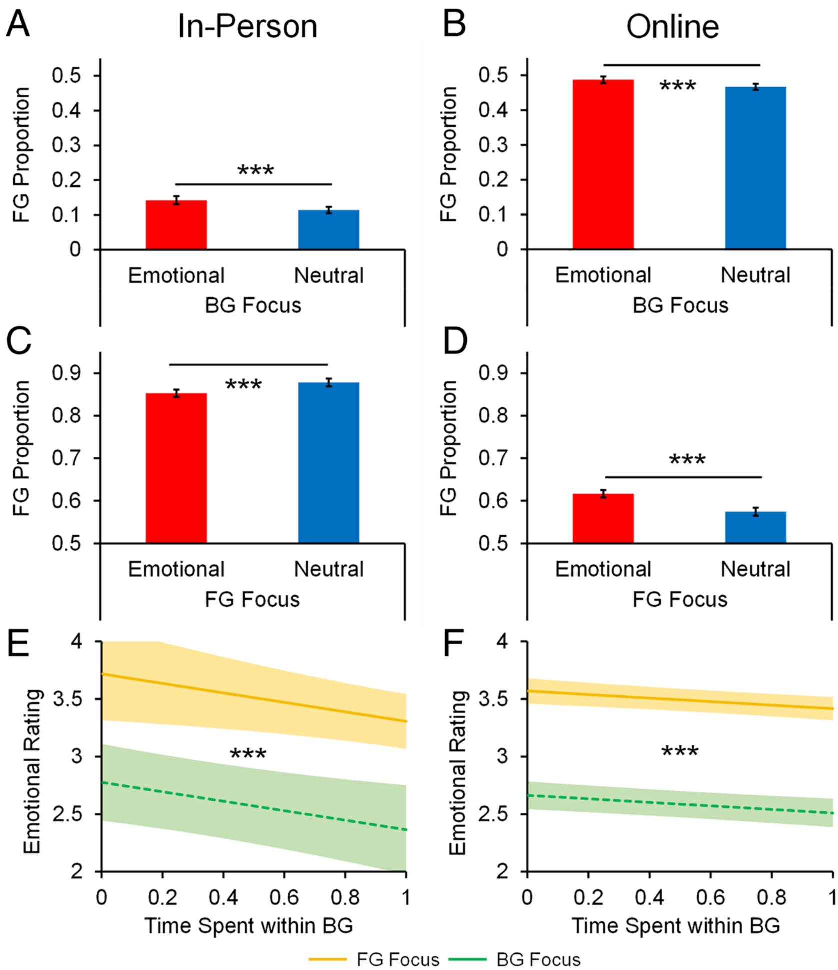


**Fig. 6** In-person vs. online comparison of the Study 2 results. **A** Representing the emotional background focus vs. neutral background focus paired *t* test on the in-person data. **B** Representing the same *t* test but on the online data, where the in-person effect replicated. **C** Representing the emotional foreground focus vs. neutral foreground focus paired *t* test on the in-person data. **D** Representing the same *t* test but on the online data, where the in-person effect did not replicate. **E** Representing multilevel regression of emotional ratings on gaze while statistically controlling for the attentional cue. **F** Representing the same regression but for the online data. Note that the *y*-axes ranges differ between the rows. *Error bars* represent 1 standard deviation above/below the mean. The *shaded region* represents a 95% confidence interval. *FG* foreground; *BG* background. ***, $p < .001$

gazed more in the foreground when this area contained emotional information. On the other hand, examination of the foreground focus trials revealed that an aversion effect was identified in person ($t[41] = -4.15$, $p < .001$; $d_z = -0.64$; Fig. 6C). That is, when asked to focus on the foreground, participants spent less time in the foreground when the area contained emotional information, relative to when it contained neutral information. However, in the online data, this analysis showed the direct opposite effect ($t[179] = 6.2477$, $p < .001$, $d_z = -0.47$; Fig. 6D). Interestingly, post hoc tests show that online participants instructed to focus on the foreground spent numerically more time (0.6%) gazing outside the emotional image entirely, compared to neutral images ($t[179] = 1.58$, $p = .12$). Although this result is not significant, the trend suggests that some attentional aversion may be occurring online, which is incompatible with the primary attentional-aversion analyses showing the total opposite. This peculiar pattern requires an explanation, and we investigated it further below using simulations.

Finally, the effect of gaze on emotional ratings, which was the main finding in the Dolcos et al. (2022) paper, was also robustly replicated. Following the procedures of the published report, this analysis was first conducted on solely the emotional trials, which yielded significant effects for both the in-person ($\beta = .11$; $p = .01$, $d_z = 0.41$; Fig. 6E) and online data ($\beta = .10$; $p < .001$, $d_z = 0.22$; Fig. 6F). In other words, emotional ratings decreased as participants spent more time in the background. The analysis was then performed using solely the neutral trials, which did not yield a significant effect for the in-person ($\beta = -.07$, $p = .32$, $d_z =$

0.16) nor the online data ($\beta = .01$, $p = .25$, $d_z = 0.02$). This demonstrates "replication" of null results, which is evidence that analyses of the online data were not "oversensitive" or tapping into spurious false-positive patterns.

### Aggregating the effect size percentages

To attain an overall estimate of the level of drop expected when transitioning from in-person to online eye-tracking research, the results from each analysis of each study were aggregated (all results listed in Table 1). As described in the Methods section, bootstrapping was used to generate a probability distribution of effect sizes for each result – analogous to how a confidence interval would be constructed. Then, these were transformed into percentages that represent the effect size preserved when transitioning from in-person to online research (rightmost column of Table 1). Aggregating these percentages revealed that the online data yielded effect sizes that were 52% [42%, 62%] of what is seen in person (Fig. 7).

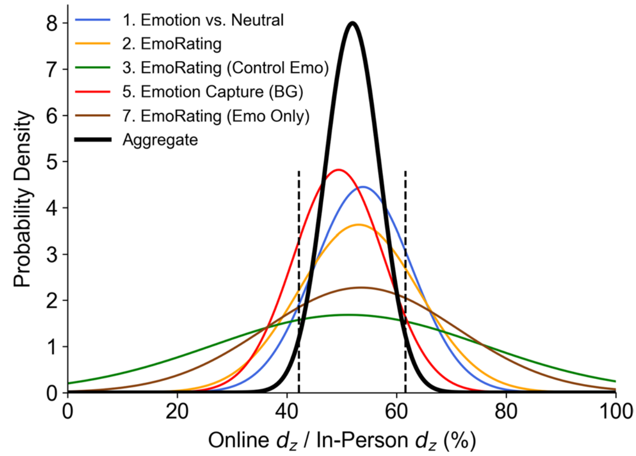


**Fig. 7** Aggregating the findings on the degree of drop expected. Aggregation was performed by taking the product of the five colored distributions (legend numbers correspond to the entries in Table 1). Note that entry 4 from Table 1 was not included, as its associated small standard deviation would cause an outsized impact on the aggregation. The aggregated distribution is shown in *black and with a thick line*. Its 95% confidence interval is indicated using *dashed lines*

### Explaining the attentional-aversion flip

To explain the flip seen between the in-person vs. online groups for the attentional-aversion result, the simulation from Study 1 was used again here. Originally, in-person participants instructed to focus on the foreground spent less time gazing toward it if it was emotional. However, adding Gaussian noise to the gaze time series causes participants to be recorded as spending significantly more time in the foreground for the emotional rather than neutral condition ($t[41] = 4.70$, $p = .001$, $d = 0.72$). Thus, the simulation reproduces the flip seen in the actual online results. Further analyses probed whether this flip can be explained by the centering bias and the degree of overlap with the center among emotional stimuli. The simulation was conducted again, now only analyzing a subset of trials where overlap with the center is equated in a pool of 30 emotional and 30 neutral stimuli. This led to the flip being undone and the original result reproduced, whereby more foreground gaze was recorded for the neutral rather than emotional condition ($t[41] = 2.23$, $p = .03$, $d_z = 0.34$). On

**Table 1** Effect-size comparisons for Studies 1 and 2

| Effect | In-Person | Online | Online / In-Person (%) |
|---|---|---|---|
| Study 1: | | | |
| 1. Emotion vs. neutral | 1.14 | 0.61 [0.41, 0.81] | 53.9% [36.3%, 71.5%] |
| 2. EmoRating | 0.99 | 0.52 [0.31, 0.74] | 53.1% [31.5%, 74.6%] |
| 3. EmoRating (Control Emo) | 0.45 | 0.23 [0.01, 0.45] | 51.2% [2.6%, 99.8%] |
| Study 2: | | | |
| 4. Attentional cue | 6.62 | 0.92 [0.82, 1.02] | 13.9% [12.3%, 15.4%] |
| 5. Emotional capture (BG) | 0.71 | 0.35 [0.22, 0.46] | 49.4% [33.2%, 65.7%] |
| 6. Emotional aversion (FG) | – 0.64 | ✕ | ✕ |
| 7. EmoRating (Emo only) | 0.41 | 0.22 [0.08, 0.36] | 53.5% [18.9%, 88.1%] |
| 8. EmoRating (Neu only) | 0.16 | 0.08 [– 0.06, 0.23] | N/A |

Each row corresponds to one of the eight analyses tested and its effect sizes (Cohen's $d_z$) for the in-person and online data sets. The rightmost column (%) represents the size of the online effect size divided by the in-person one. "✕" signifies that the effect did not emerge in a given data set. Concerning the last row, the "N/A" indicates that the comparison of effects sizes is not applicable, given that no significant effect of gaze on emotional ratings was found among the neutral trials for either data set (i.e., the in-person null result was "replicated" in the online data). *FG* foreground, *BG* background, *Emo* emotional, *Neu* neutral the other hand, the simulation can also be conducted while maximizing discrepancies in center overlap. This causes the flip to be amplified ($t$[41] = 7.00, $p < .001$, $d_z = 1.08$) relative to the original flip ($d_z = 0.72$). Hence, the attentional flip is rooted in the centrality of the foregrounds of individual stimuli.

## Discussion

Altogether, the two studies provide insight into how the spatial error of online eye tracking impacts analyses of psychological effects on gaze. The two studies showed that seven of the original eight in-person results were replicated online, and that researchers can expect effect sizes 52% [42%, 62%] of those seen in person. Additionally, the studies demonstrated a centering bias that arises during online eye-tracking research. The bias explains why one result did not replicate online, where it even showed a significant flip in the opposite direction of what was seen in-person. Altogether, our results provide a quantification of how much weaker online eye-tracking data can be expected to be in replicating in-person results. Furthermore, we also provide evidence for a "centering bias" in the online eye-tracking data that explains why one of our in-person findings did not replicate in the online data.

## Replications and effect sizes

The present research builds on the growing online eye-tracking literature (Schneegans et al., 2021; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021; Slim & Harsuiker, 2022; Vos et al., 2022; Degen et al., 2021). To varying degrees, these earlier studies arrived at the same conclusion: although data quality is reduced, online eye tracking partially replicates patterns found by in-person eye tracking. These earlier studies also provided evidence on several important methodological points (e.g., on the effects of glasses and head positioning) and yielded initial estimates of data quality outside the context of psychological tasks and effects (e.g., the distance from gaze measurement to fixation cross). The present research advances this body of work by quantifying the quality of online eye-tracking data with respect to detecting psychological effects within experiments. Hence, the current studies inform the feasibility and usefulness of adopting this new technology by researchers.

The present studies were best suited to expand the extant research for several reasons. First, our studies used the same types of participants for both in-person and online samples, while earlier research collected in-person and online data from different populations (local community vs. online crowdsourcing), which may create potential confounds during comparison (Ahler et al., 2019). Second, unlike earlier works, which attempted to replicate between one and three hypotheses, the current research targeted eight hypotheses in total across two studies, which covered a diverse range of effect size magnitudes. These aspects increase the generalizability of the conclusions. Third, the present research uniquely targeted highly reliable emotion processing effects (Carretié, 2014), which were expected to arise both in person and online, and would be less susceptible to other sources of in-person vs. online differences (e.g., participant engagement; Ahler et al., 2019).

Ultimately, the present results suggest that online eye tracking will yield effects sizes 42–62% of what is found in person. Additionally, online eye-tracking will lead to the time recorded within a given interest area to approach chance rates (e.g., 50% gaze for each condition), mirroring the results by Schneegans et al. (2021). Regarding the effect size drops, we do not wish to understate that they are substantial. However, they are surmountable. For example, to achieve appropriate statistical power (power = 80%, two-tailed $\alpha = .05$), an in-person study targeting a moderate effect ($d_z = 0.50$) requires 34 participants, while an online study ($d_z = 0.25$; 50% drop) would require roughly 128 participants. Researchers will inevitably also need to recruit further participants beyond that, as this sample size does not account for issues related to non-compliance or other challenges that arise with online research. For instance, we needed to exclude 18% of participants' data because of low response rates or unusable gaze data. Nonetheless, even with the need for exclusion, these sample sizes remains attainable, given the ease of online data collection and the trend in psychology for larger samples (Sassenberg & Ditrich, 2019).

Although excluding 18% of online participants is a high rate compared to most psychological studies, it is notably lower than typical in online eye-tracking research (e.g., Schneegans et al., 2021; Slim & Hartsuiker, 2022; Yang & Krajbich, 2021). As part of preliminary analyses, we attempted more conservative exclusion procedures to clean the data, such as excluding participants who self-reported high head movement or who wore glasses. These cleaning procedures enhanced online effect sizes, but the benefits did not compensate for the reduction in sample size (Cohen's $d$ increased while $t$ values decreased). Hence, we opted to include those data sets.

One outstanding question for interpreting our results is the extent to which the effect size drops should be understood as technological differences between infrared-based vs. webcam-based eye tracking or due to behavioral differences. Several previous studies have demonstrated that online crowdsourced samples are of lower quality (e.g., on Amazon Mechanical Turk; Kees et al., 2017). Our research used participants samples of the same type (college students) for both the in-person and online conditions and thus provides a cleaner comparison than studies comparing college students to crowdsourced samples. However, even with the same population, data quality degrades when transitioning online. In the present studies, we saw nominal patterns suggesting that online participants had dampened emotional reactions and exerted less attentional control, possibly reflecting lower engagement. This likely explains some differences between the in-person and online data. For instance, although the centering bias goes a long way in explaining the flip seen in the Study 2 attentional-aversion result, the dampened emotional processing by online participants likely added to it. The aversion effect involves arousal overriding participants' top-down attentional control, but if participants' affective responses are weak and they do not exert substantial control in either condition, the impact of emotional stimuli will be smaller. Hence, aside from just the impact of changing technologies, when transitioning to online studies, researchers must also consider possible behavioral changes, which may impact some hypotheses more than others.

## Recommendations regarding the centering bias in online data

Online eye tracking's high spatial error (3–4.5° visual angle) is well known (Papoutsaki et al., 2016; Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022), but how this error impacts researcher's abilities to detect psychological effects has been unclear. Our findings bridge these levels of analysis and come with specific recommendations: Authors must be careful in designing their stimuli to avoid interference due to the centering bias, such as by having symmetry in their interest areas. For example, in decision-making or memory tasks, choices are often represented as rectangles, which are the interest areas. When two choices are available, placing them left/right or top/down avoids centering bias interference (Schneegans et al., 2021; Yang & Krajbich, 2021), but when three or more choices are available, risks arise if some choices are closer to the center (Hutt et al., 2023). Such dangers can notably be mitigated via counterbalancing or analyzing the different interest area locations independently. These good practices, already common for in-person research, become more important when transitioning to online eye-tracking collection and are also particularly relevant when using complex pictorial stimuli.

Counterbalancing is more challenging for designs like the present one, which used realistic stimuli and interest areas with complex shapes. In these cases, we advise researchers to perform quantifications like ours, measuring the extent to which the stimuli's interest areas are susceptible to the centering bias. Beyond just the centering bias we found, related risks may also exist that future studies can likewise see via simulations (e.g., bias related to interest areas' shapes). Although the present report's partial focus on effect sizes may seem to frame online eye tracking as simply uniformly lower quality than in-person eye tracking, this would be an inappropriate simplification, as researchers must be cognizant of how noise may interact with their own specific interest areas and comparisons.

## Broad implications

Critically, we would like to also emphasize that implementing online eye tracking is surprisingly easy within flexible experiment platforms (e.g., https://pavlovia.org/). By relying on the freely available *WebGazer.js* package, we were able to incorporate eye tracking within our online study in just a couple of hours and two dozen lines of code. Hence, the current results do not speak just to labs already conducting eye-tracking research. Given the low "cost" of online eye tracking, it is worthwhile for a wide range of research to incorporate eye tracking, even in cases where gaze is not the primary focus or just for exploratory purposes. For example, in emotional memory research, the attention-capturing effect of emotional stimuli is often a potential confound that muddles investigation of other mechanisms by which emotion impacts memory (Bogdan et al., 2023; Riggs et al., 2011; Voss et al., 2017). Likewise, the effect of aging on memory too involves perceptual mechanisms/confounds, as aging changes people's viewing patterns, which in turn impact encoding (Voss et al., 2017). Beyond memory studies, decision-making research also stands to benefit from incorporating eye tracking – e.g., investigating the time participants spend gazing towards each available choice is informative for understanding how they arrive at their final decision (Brunyé & Gardony, 2017; Fiedler & Glöckner, 2012; Krajbich et al., 2010; Kwak et al., 2015). Thus, the greatest benefit of webcam-based eye tracking may not come from vision labs taking their in-person experiments online. Rather, its inclusion may benefit other areas of research incorporating eye tracking to investigate/account for the role of attention in various psychological phenomena. Although online eye tracking is far from perfect, it can at least provide important evidence on the role of attentional processes in research designs involving visual stimuli.

## Caveats

Although Study 1 identified repeated patterns of strong replication (three out of three), Study 2 did not show quite the same high level of consistency (four out of five).

Additionally, for Study 2, one of the analyses replicated but with a very large drop in the effect size. Although this was by far the largest drop across either study, it should be noted that the online effect size remained large, in an absolute sense ($d_z = 0.9$), and the in-person effect was far beyond what is typically studied in psychological research ($d_z = 6.6$). Given these large magnitudes, this analysis may be less representative and does not change the conclusion of the investigation. Finally, concerning the result which did not replicate at all, it is worth noting that it was among the more subtle in-person effects (see the numerically small in-person differences in-person in Fig. 6C). Nonetheless, the main finding by Dolcos et al. (2022) was replicated with a robust online effect size. This final result is probably the most representative, as the in-person effect size is most consistent with what is typically targeted by psychological research (medium and small-to-medium effects; Gignac & Szodorai, 2016; Lovakov & Agadullina, 2021).

## Conclusion

In sum, the present results suggest that effective online eye-tracking research is currently possible. Across two studies and eight different analyses, seven of the patterns found in person were replicated in the online eye-tracking data. Further, effect size comparison between the two data sets showed that well-powered online eye-tracking research is possible with realistic sample sizes. Hence, researchers studying gaze in-person can perform some of their studies online, and researchers already conducting online research can benefit from recording eye movements. Although online eye tracking also comes with limitations, given the relative ease of online collection and its wide-reaching utility (e.g., for emotion, memory, and decision-making research), researchers should begin to consider this emerging technology.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02143-z.

**Acknowledgments** This research was carried out in part at the University of Illinois' Beckman Institute for Advanced Science & Technology. During the preparation of this manuscript, P.C.B. was supported by a Predoctoral Fellowship provided by the Beckman Foundation and a Dissertation Completion Fellowship provided by the University of Illinois, and F.D. was supported by an Emanuel Donchin Professorial Scholarship in Psychology from the University of Illinois. The authors thank Margaret O'Brien, Anna Madison, and Chen Shen for their assistance with data collection for the in-person versions of the tasks. The authors also thank Dolcos Lab members for their help with stimulus creation.

**Code availability** The analysis code has been deposited in a public GitHub repository, alongside the data (https://github.com/paulcbogdan/Suitability_Online_ET).

**Authors' contributions** F.D. and S.D. conceived of the tasks; P.C.B. implemented the online versions of the task and the online eye tracking with guidance from F.D. and S.D.; P.C.B. collected the online data; P.C.B. designed the analytic approach with feedback from F.D., S.D., S.B., and A.L.; P.C.B. performed the analyses; P.C.B. wrote the first draft of the manuscript and revised it based on feedback from F.D, S.B., A.L., and S.D. All authors approved of the final submission.

**Funding** This research was supported by research funds from the University of Illinois to F.D. and S.D.

## Declarations

**Conflict of interest** The authors declare no conflicts of interest.

**Ethics approval** Protocols performed by the current research were approved by the University of Illinois Institutional Review Board.

**Consent to participate** Informed consent was obtained from all individual participants included in the study.

**Consent for publication** Not applicable.

## References

Ahler, D. J., Roush, C. E., & Sood, G. (2019). *The micro-task market for lemons: Data quality on Amazon's mechanical Turk* Paper presented at the Meeting of the Midwest Political Science Association.

Amasino, D. R., Sullivan, N. J., Kranton, R. E., & Huettel, S. A. (2019). Amount and time exert independent influences on intertemporal choice. *Nature Human Behaviour, 3*(4), 383–392.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv*. https://doi.org/10.18637/jss.v067.i01

Bogdan, P. C., Dolcos, F., Katsumi, Y., Dolcos, S., O'Brien, M., Iordan, A. D., et al. (2023). *Reconciling opposing effects of emotion on relational memory: Behavioral, eye-tracking, & brain imaging investigations* Submitted.

Brunyé, T. T., & Gardony, A. L. (2017). Eye tracking measures of uncertainty during perceptual decision making. *International Journal of Psychophysiology, 120*, 60–68.

Carretié, L. (2014). Exogenous (automatic) attention to emotional stimuli: A review. *Cognitive, Affective, & Behavioral Neuroscience, 14*(4), 1228–1258.

Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods, 43*(2), 468–477. https://doi.org/10.3758/s13428-011-0064-1

Degen, J., Kursat, L., & Leigh, D. D. (2021). *Seeing is believing: Testing an explicit linking assumption for visual world eye-tracking in psycholinguistics* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Dolcos, F., Bogdan, P. C., O'Brien, M., Iordan, A. D., Madison, A., Buetti, S., & Dolcos, S. (2022). The impact of focused attention on emotional evaluation: An eye-tracking investigation. *Emotion, 22*(5), 1088–1099.

Dolcos, F., Katsumi, Y., Bogdan, P. C., Shen, C., Jun, S., Buetti, S., & Dolcos, S. (2020a). The impact of focused attention on subsequent emotional recollection: A functional MRI investigation. *Neuropsychologia, 138*, 107338.

Dolcos, F., Katsumi, Y., Moore, M., Berggren, N., de Gelder, B., Derakshan, N., & Dolcos, S. (2020b). Neural correlates of emotion–attention interactions: From perception, learning, and memory to social cognition, individual differences, and training interventions. *Neuroscience & Biobehavioral Reviews, 108*, 559–601.

Dolcos, F., Katsumi, Y., Shen, C., Bogdan, P. C., Jun, S., Larsen, R., & Dolcos, S. (2020c). The impact of focused attention on emotional experience: A functional MRI investigation. *Cognitive, Affective, & Behavioral Neuroscience, 20*(5), 1011–1026.

Ehinger, B. V., Groß, K., Ibs, I., & König, P. (2019). A new comprehensive eye-tracking test battery concurrently evaluating the pupil labs glasses and the EyeLink 1000. *PeerJ, 7*, e7086.

Fiedler, S., & Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology, 3*, 335.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78.

Goodman, A. M., Katz, J. S., & Dretsch, M. N. (2016). Military affective picture system (MAPS): A new emotion-based stimuli set for assessing emotional processing in military populations. *Journal of Behavior Therapy and Experimental Psychiatry, 50*, 152–161. https://doi.org/10.1016/j.jbtep.2015.07.006

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*(2), 348–362. https://doi.org/10.1037/0022-3514.85.2.348

Hutt, S., Wong, A., Papoutsaki, A., Baker, R. S., Gold, J. I., & Mills, C. (2023). Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior Research Methods*, 1–17.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's mechanical Turk. *Journal of Advertising, 46*(1), 141–155.

Kragel, J. E., & Voss, J. L. (2022). Looking for the neural basis of memory. *Trends in Cognitive Sciences, 26*(1), 53–65.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13*(10), 1292–1298.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Kwak, Y., Payne, J. W., Cohen, A. L., & Huettel, S. A. (2015). The rational adolescent: Strategic information processing during decision making revealed by eye tracking. *Cognitive Development, 36*, 20–30.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6*(1), 1–11.

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology.*

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Marchewka, A., Zurawski, L., Jednorog, K., & Grabowska, A. (2014). The Nencki affective picture system (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods, 46*(2), 596–610. https://doi.org/10.3758/s13428-013-0379-1

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*(3), 466.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, J., Huang, J., & Hays, J. (2016). *WebGazer: Scalable webcam eye tracking using user interactions*. In *Paper presented at the IJCAI-16: International joint conference on artificial intelligence*. USA https://par.nsf.gov/biblio/10024076

R Core Team(2013) R: A language and environment for statistical computing.

Riggs, L., McQuiggan, D. A., Farb, N., Anderson, A. K., & Ryan, J. D. (2011). The role of overt attention in emotion-modulated memory. *Emotion, 11*(4), 776.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science, 2*(2), 107–114.

Schneegans, T., Bachman, M. D., Huettel, S. A., & Heekeren, H. (2021). *Exploring the potential of online webcam-based eye tracking in decision-making research and influence factors on data quality*.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods, 50*(2), 451–465.

Slim, M. S., & Hartsuiker, R. J. (2022). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer. Js. *Behavior Research Methods*, 1–19.

Strauss, G. P., Ossenfort, K. L., & Whearty, K. M. (2016). Reappraisal and distraction emotion regulation strategies are associated with distinct patterns of visual attention and differing levels of cognitive demand. *PLoS One, 11*(11), e0162290.

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., & Reiss, K. M. (2020). Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics, 104*, 147–200.

Vos, M., Minor, S., & Ramchand, G. C. (2022). *Comparing infrared and webcam eye tracking in the visual world paradigm*.

Voss, J. L., Bridge, D. J., Cohen, N. J., & Walker, J. A. (2017). A closer look at the hippocampus and memory. *Trends in Cognitive Sciences, 21*(8), 577–588.

Wessa, M., Kanske, P., Neumeister, P., Bode, K., Heissler, J., & Schönfelder, S. (2010). EmoPicS: Subjective and psychophysiological evaluation of new imagery for clinical biopsychological research. *Z. Klin. Psychol. Psychother. Suppl, 1*, 11–77.

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision making, 16*(6).


**Open Practices Statement** The data and analysis code will be deposited in a publicly available repository, pending IRB approval and upon article acceptance. The experiments were not pre-registered.