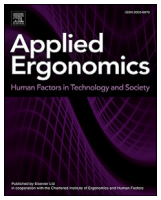
Applied Ergonomics 117 (2024) 104242



Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo



Comparison of two pupillometric measures as indicators of cognitive strain and under the influence of screen luminance

Nikolas J. Schierhorst [a,*], Vera B. Rick [a], Julia N. Czerniak, Christopher Brandl [a,b], Alexander Mertens [a], Verena Nitsch [a,b]

[a] Institute of Industrial Engineering and Ergonomics, RWTH Aachen University, Germanny
[b] Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany

ARTICLE INFO

Keywords:
Index of Cognitive Activity (ICA)
Index of Pupillary Activity (IPA)
Pupillary light reflexes

ABSTRACT

The pupil diameter has been shown to provide insight to a person's experienced cognitive strain. Pupillary light responses, however, make this measure unreliable in uncontrolled settings. Two derived indicators—Index of Cognitive Activity (ICA) and Index of Pupillary Activity (IPA)—aim to 'eliminate' lighting influences, changing based only on the perceived cognitive strain. The IPA potentially offers a valuable alternative to the ICA through its fully transparent calculation, which lifts the restrictions to proprietary software and supported eye trackers. The measures are examined and compared based on two experimental studies; (i) as indicators of cognitive strain during mental arithmetic tasks and (ii) under different conditions of computer screen luminance. Results indicate that neither indicator differentiates between the increasing levels of cognitive strain. Differences in screen luminance are reflected in both indicators, although differently between the conditions. Both results contradict the claims of the indicators and further investigations are thus required.

1. Introduction

Detecting, quantifying, and assessing cognitive strain[1] is a vital step for the (cognitive) ergonomic advancement of current and future workplaces. An effective and reliable measurement of cognitive strain could benefit many areas. For one, the cognitive strain experienced by workers can be effectively managed to avoid an overworked or overstrained workforce, to not only ensure adequate worker safety, but also help to reduce human errors due to inattentiveness or fatigue (Young et al., 2015). Other application areas could include the monitoring of worker's cognitive strain during training tasks, for example by adapting task type or difficulty to correspond with the worker's skill and knowledge level as reflected in their experienced cognitive strain.

Various indicators based on the human eyes offer insight to a person's cognitive strain. Next to the measurement of, for example, blink rates or saccades, the diameter of the pupil has been shown to reflect mental activity (Armougum et al., 2020; Marquart et al., 2015). In general, research indicates that the pupil dilates with increasing cognitive strain (Charles and Nixon, 2019; van der Wel and van Steenbergen, 2018). A major drawback with the use of the pupil diameter, however, is the pupil's reaction to differing light influences (Beatty and Lucero-Wagoner, 2000). Changing light conditions evoke pupillary responses which interfere with the reliability of this indicator, especially in field applications where such changes cannot be eliminated or accounted for. To address this shortcoming, the Index of Cognitive Activity (ICA) and the Index of Pupillary Activity (IPA) were introduced (Marshall, 2000; Duchowski et al., 2018).

The aim of this work is to examine and compare both indicators. This includes an assessment of their suitability to detect and measure cognitive strain. Their supposed benefit over the pupil diameter—the absence of lighting influences—is also examined. In the context of this work, cognitive strain is induced by the solving of mental arithmetic tasks whilst lighting influences are based on differing conditions of (computer) screen luminance.

1.1. Pupillometric indicators of cognitive strain

The pupil diameter is an often examined physiological measure of cognitive strain induced by a variety of different tasks (Hess and Polt, 1964; Kahneman and Beatty, 1966; Ahern and Beatty, 1979; Beatty,

* Corresponding author.
E-mail address: n.schierhorst@iaw.rwth-aachen.de (N.J. Schierhorst).

[1] A definition of the term *cognitive strain* used in this work can be found in the appendix, p. 8.

https://doi.org/10.1016/j.apergo.2024.104242
Received 21 March 2023; Received in revised form 22 January 2024; Accepted 23 January 2024
Available online 2 February 2024
0003-6870/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1982). An advantage over other physiological measures includes its relatively quick reaction to stimuli, which allows for a momentary assessment of the experienced cognitive strain (Laeng and Alnaes, 2019). Moreover, it benefits from aspects of the eye tracking systems used to record the measure and the large variability of the setup of such systems (Holmqvist, 2017): Depending on the use case, eye tracking systems can either be worn by the person directly or placed in a set place and used remotely. In either case, they are completely non-invasive and often barely noticeable by the measured person and thus have little to no impact on the task itself (Duchowski, 2017).

The ICA is an indicator of cognitive strain and is derived from the pupil diameter (Marshall, 2000). It does, however, differ in one key aspect which gives it a major advantage over the pupil diameter, especially in less controlled environments, outside of laboratory settings: through decomposition of the pupil diameter signal by means of wavelet analysis, the pupillary changes caused by light can be effectively 'filtered out'. The resulting signal is thus indicative of the cognitive strain without influences of differing lighting conditions. This is achieved by separating the pupil diameter signal into a tonic and phasic component. This is grounded in the fact that an increase in cognitive activity is associated with an increase in the phasic component, whilst the tonic component remains relatively stable (Peysakhovich et al., 2015). The decomposition is based on Daubechies wavelets and a minimax threshold estimation. The resulting ICA signal is given in a sampling rate of 1 Hz, showing the ratio of phasic measurements, excluding all measurement errors, over all measurements in the given timeframe of one second. The resulting values range from 0 to 1, with higher values indicating a higher level of cognitive strain (EyeTracking, 2014).

Since its introduction, however, adoption has been relatively limited and the indicator is subject of only a few published studies. A recently published meta-analysis, for example, found only 14 relevant studies in 13 publications, only four of which also examined lighting influences (Czerniak et al., 2021a).

A factor which presents a hurdle to adoption and is potentially one of the reasons for the limited application, is the closed-source nature of the ICA. This presents possible financial restrictions associated with the acquisition of expensive commercial software which is required for the calculation and further restricts the used hardware to supported systems (EyeTracking, 2022, 2020).

The IPA is inspired by the ICA and builds upon the same principles of 'filtering out' the influences of light using wavelet-transforms (Duchowski et al., 2018). The resulting values also range from 0 to 1, with greater values reflecting higher cognitive strain. The authors state the main differences to be the *"choice of wavelet, use of the modulus maxima, and a different thresholding approach"* (Duchowski et al., 2018). Specifically, a modified version of the Debauchies walvet—symlet wavelets—are used, whereby the size of the wavelet is adjusted based on the eye tracker sampling rate. In place of a minimax thresholding, Duchowski et al. (2018) use the local maxima of the wavelet modulus to detect sharp changes in the pupil diameter signal. The coefficients are thresholded based on "universal thresholding" which is based on the standard deviation of the noise (Duchowski et al., 2018).

Furthermore, in contrast to the ICA, the calculation of the IPA is fully transparent and the code completely open source, which offers many benefits. Since the calculation is done using python scripts and the raw pupil diameter signal, no additional software is required. The complete public documentation of the IPA thus has the benefit of a much higher availability and accessibility. Unrestricted insight to the code and its workings also leave the calculation method open to review and audit, which could lead to modifications and potential improvements. This could prove valuable, especially considering both indicators' somewhat inconclusiveness regarding their ability to measure cognitive strain and their robustness towards lighting influences. A non-exhaustive list of previously published results is presented below. A more complete overview of the literature to the ICA can be found in the meta-analysis by Czerniak et al. (2021a).

Supporting literature Marshall (2007) presents three studies in which the ICA was used to successfully measure cognitive strain in different tasks; solving arithmetic (addition, subtraction, and multiplication) problems, driving, and during a visual search task. Furthermore, the ICA was shown to be robust towards ambient lighting influences in the first study including the arithmetic problems, which the participants solved in both light and dark lighting conditions. Using a combination of smooth pursuit, mental arithmetic, and attention tasks, Bartels and Marshall (2012) again showed significant results for the ICA to differentiate between different levels of cognitive strain. Furthermore, the comparison across four different eye tracking systems showed no influence of the hardware on the reliability of the measure. Using a n-back task as a cognitive stimulus, Devos et al. (2020) showed the ICA to be a reliable measure of cognitive activity in older adults. A second study using the n-back tasks by Fairclough et al. (2009) showed differences between the varying difficulty levels, although the results were not statistically significant.

In the introductory publication to the IPA, Duchowski et al. (2018) examined the indicator's ability to detect different levels of cognitive strain induced by a digit-span test—similar to the one used by Marshall et al. (2004)—to validate the ICA. The results showed significant differences between the different task difficulties. Later examinations by Weber et al. (2021) showed the same result of differentiation between different task difficulties in memory span tasks. Gunawardena et al. (2019) examined the IPA outside of a laboratory setting for the measurement of a surgeon's cognitive strain during surgery. Measurements during the carrying out of surgical procedures showed that the indicator was able to differentiate between novice and expert surgeons—a difference in skill level which was also reflected in the performance measure.

Contradicting literature Whilst most publications seem to support the indicators' abilities to measure cognitive strain while filtering out lighting influences, some contradicting results exist. For the ICA, this applies to its ability to detect cognitive strain as well as its robustness towards lighting influences. For example, the two studies carried out by Rerhaye et al. (2018), two types of tasks—a spatial task and a Stroop task—were examined in different lighting conditions. While it was possible to differentiate between different task difficulties of a spatial task, the same was not true when participants were presented with a Stroop task. The lighting conditions were also shown to influence the ICA during the Stroop task, but not during the spatial task, although conditions were identical. Opposing results were presented by Kahya et al. (2018), which, while not explicitly examining the effects of lighting conditions on the ICA, included an experimental condition in which the participant's eyes were occluded, effectively changing the ambient luminance and thus leading to a pupillary response which was also reflected in the ICA. Using reading tasks as the cognitive stimulus, Debue and van de Leemput (2014) showed non-significant results for the ICA as an indicator of cognitive strain. Korbach et al. (2018) report similar results in learning tasks.

Much fewer studies exist for the IPA. Książek et al. (2021), for example, examined the indicator in the context of linguistic processing using a speech-in-noise test, in which participants had to repeat previously heard sentences. Results could not confirm a differentiation of different cognitive states based on the IPA. In the same study, two different ambient lighting conditions—light and dark—were also examined. Here, the indicator's claims could not be confirmed, as the luminance conditions had a significant effect on the IPA (Książek et al., 2021).

1.2. Research questions

Based on the considerations above, we examine if and to what extent the IPA can offer an alternative to the ICA. Thus, the following research questions are defined:

Research Question 1 *(RQ 1)*. In which ways does the IPA differ from the ICA in its sensitivity towards different levels of cognitive strain?

Research Question 2 *(RQ 2)*. In which ways does the IPA differ from the ICA in its robustness towards different conditions of screen luminance?

2. Method

The following work examines both indicators in the two relevant aspects: regarding their ability to detect and measure cognitive strain and regarding their robustness towards lighting influences. The analysis itself is a secondary analysis of previously carried out investigations of the ICA by Czerniak et al. (2021b), differing in certain aspects and extended by the calculation and examination of the IPA. To allow for a direct comparison of the indicators, analyses of both indicators were carried out independently of the previous publications.

Based on the research questions identified and presented above, the hypotheses are formulated following the claims of the respective indicators; see Marshall (2000); Duchowski et al. (2018). First, to address RQ 1, the hypotheses state that *there is a statistically significant difference in cognitive strain as indicated by (a) the ICA and (b) the IPA between different levels of task difficulty*. Both indicators follow a positive scale and relationship with the associated cognitive strain. As such, higher levels of cognitive strain should be reflected in higher values of both indicators. Regarding the robustness of the indicators towards different lighting conditions, the focus of this work lies on the effects of the screen luminance, as stated in RQ 2. Although the claims of the indicators suggest that no difference is found between different luminance conditions the hypotheses are, in a first step, set up to comply with traditional comparative hypotheses testing (Walker and Nowacki, 2011). The hypotheses to be examined thus counter the claims of the indicators to state that *there is a statistically significant difference in (a) the ICA and (b) the IPA between bright and dark conditions of screen luminance*. In the case of non-significant test results an additional equivalence test is performed, as a non-significant result is not indicative of the absence of an effect, which is often overlooked (Lakens, 2017).

2.1. Study I: As indicators of cognitive strain

Experimental design & procedure The first study by Czerniak et al. (2021b) measured the ICA of participants while they were given the task of solving multiplication problems displayed centrally on a computer screen. Problems were solved mentally, without the use of calculation aids. Overall the experimental conditions were divided into five difficulty levels, presented in order of increasing difficulty. The easiest condition included multiplication problems, such as 6×7, while the final condition contained much more difficult problems, such as 36×37. A table of all possible multiplicands can be found in Table A.1 in the Appendix. For each condition participants were given a total of 120 seconds to answer as many problems as possible, with a maximum of 22 problems. Participants were instructed to keep their gaze towards the computer screen during the solving of the problems. Answers were given orally and noted by the observer, who gave no immediate feedback to the correctness of the answers. In addition to the pupil diameter and ICA values collected by the eye tracking system, performance of the primary task was recorded. Performance was measured with both the total number and the number of correctly solved problems per condition. Between the conditions, participants were asked to complete the NASA-TLX questionnaire (Hart and Staveland, 1988) to assess their subjective strain and given a break before the onset of the next task. The NASA-TLX was carried out in its entirety, including the pairwise comparison of six items—*mental demand, physical demand, temporal demand, effort, performance, and frustration*—to determine the respective weighting. The resulting NASA-TLX score is a weighted sum of the six items, ranging from 0 to 100, with higher values indicating higher strain. Before beginning the study, participants were thoroughly informed about the study's purpose, the anonymized use and storage of the data, and their rights, including the option to withdraw without consequences.

Apparatus Measurement of the pupil diameter was done using a remote eye tracking system; the FOVIO FX3 from EyeTracking (2021), placed directly beneath the computer screen on which the arithmetic tasks were displayed. Participants were seated in front of a height adjustable desk. Sampling rate for the pupil diameter measurements was 60 Hz. The bundled EyeWorks software package was used to create individual task conditions (EyeTracking, 2022). The experiment itself was carried out in a laboratory at the Institute of Industrial Engineering and Ergonomics at RWTH Aachen University. Ambient lighting conditions were kept constant throughout the experimental procedures.

Participants The study examined a total of 22 participants between the ages of 20 and 56 (AM = 28.4, SD = 8.8), which were recruited from RWTH Aachen University through flyers or personal contact on the campus. Of the participants, 14 identified as female and eight as male. Exclusion criterion was a visual impairment which could not be compensated by a corrective aid (e.g., glasses or contact lenses). This was tested beforehand using a Vistec Rodatest device. If a corrective aid was required, it was worn during the experiment.

Statistical analysis & hypothesis test To examine the effects of the tasks and thereby induced cognitive strain, a repeated measures ANOVA was conducted to explore the differences between the task conditions. The assumption of normality was tested using the Shapiro-Wilk-Test (Shapiro and Wilk, 1965), which showed that in all but a few cases the values follow a normal distribution. The ANOVA, however, is generally robust to minor violations hereof and is thus not dismissed (Bortz and Schuster, 2010). Sphericity of the data, was tested using Mauchly's Test of Sphericity (Mauchly, 1940) and corrected according to Greenhouse and Geisser (1959). Effect sizes were calculated as partial η^2. The few number of outliers[2] present in the dataset, were not removed to maintain a complete population sample and because they were not found to significantly change the statistical results.

2.2. Study II: Under influence of screen luminance

Experimental design & procedure The second study by Czerniak et al. (2021b) measured the ICA of participants under different effects of perception. Participants were given no other task than concentrating their gaze onto the computer screen placed in front of them. Following a baseline measurement in front of a black background, the screen automatically cycled between six different images, either with a black or white background for a total of twelve different conditions.[3] Each image was shown once for a total of 30 seconds. The order hereby was randomised across all participants using a Latin square to ensure that no two participants had the same sequence of images. As the aim of the study was to examine perceptional effects, the conditions included images with certain information, such as a centred fixation cross, dot matrices, or moving objects (Czerniak et al., 2021b). Two conditions were plain backgrounds, one in white and one in black colour, without any additional information. These were the only conditions used in the presented analysis as the conditions of bright and dark screen luminance, respectively. During these conditions, no additional task was presented. Instructions were to only keep the gaze fixated towards the center of the computer screen. A fixation cross was presented before the condition, however removed for the bright and dark screen luminance conditions. The following procedure and handling of the recorded data was identical to Study I (Section 2.1).

[2] Defined based on the interquartile range (Bortz and Schuster, 2010, p. 32).
[3] A depiction of all conditions can be found in Czerniak et al. (2021b).

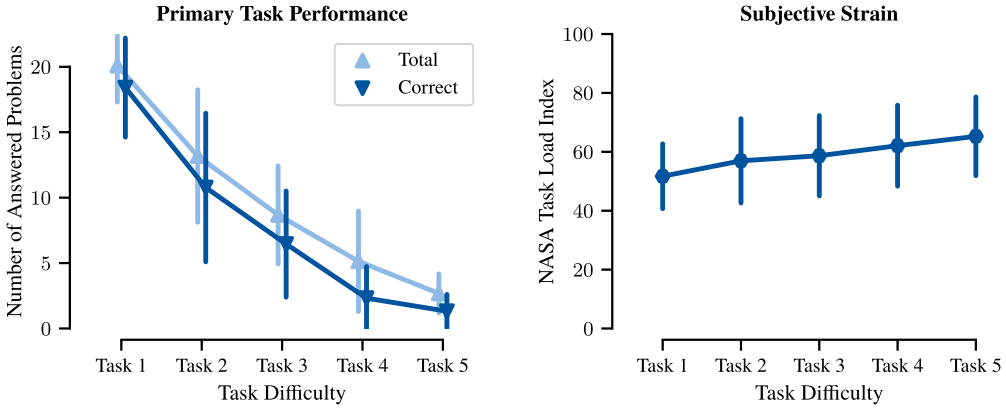


Fig. 1. Task performance and NASA-TLX values (AM ± SD) for different task difficulty conditions. Adapted from Czerniak et al. (2021b).

Apparatus The experimental setup was identical to Study I (Section 2.1), with the addition of a chin rest to stabilise the participants' gaze.

Participants The study examined a total of 20 participants between the ages of 19 and 34 (AM = 25.9, SD = 4.5), which were recruited from RWTH Aachen University through flyers or personal contact on the campus. Of the participants, eleven identified as male and nine as female. The exclusion criterion regarding visual acuity and testing hereof was identical to Study I (Section 2.1). The calculated ICA values for two participants, however, showed a high number of measurement errors in at least one of the conditions and were thus removed for the analysis, reducing the total sample size to $N_{ICA} = 18$.

Statistical analysis & hypotheses test The statistical analysis of the two conditions of dark and bright luminance is done using the Wilcoxon-Test for paired samples (Wilcoxon, 1945). The non-parametric test was chosen due to violations of the normality assumption, especially for the IPA (Bortz and Schuster, 2010). Effect sizes were calculated accordingly as r, as the Z-statistic divided by the square root of the sample size. As described in the analysis to Study I, the few present outliers were not removed for the following analysis.

3. Results

3.1. ICA and IPA as indicators of cognitive strain

Fig. 1 shows the mean performance and NASA-TLX values across the different difficulty conditions. It can be observed that with increasing task difficulty the performance—as measured by the number of answered problems in the given time—decreased. While most participants answered all 22 multiplication problems in the first condition (AM = 20, SD = 3), by the third condition the average number of answered problems dropped to less than half (AM = 9, SD = 4). In the final, most difficult condition, the number of answered problems dropped to three (AM = 3, SD = 1). The number of correctly solved problems closely followed this trend. In the first condition, most problems were solved correctly (AM = 18, SD = 4). In the third condition this number dropped to an average of six (SD = 4), and to almost no correctly solved problems in the final task condition (AM = 1, SD = 1). The subjective experienced strain as measured by the NASA-TLX increased with task difficulty. The first condition was rated with the lowest strain (AM = 51.7, SD = 10.8) and increased steadily over the course of the experiment to the final condition (AM = 65.1, SD = 13.1). The individually weighted NASA-TLX items show that the increase in subjective strain was mainly due to the steady increase of the *mental demand* and *effort* dimensions, while the other items remained relatively stable across the task difficulty levels. An overview of these items across all task difficulty levels can be found in Fig. A.1 in the Appendix. These effects were also confirmed by the ANOVA presented in the original publication, which showed significant effects, confirming a successful manipulation of the conditions with increasing levels of cognitive strain (Czerniak et al., 2021b).

Table 1
ANOVA Test Statistics.

Indicator*	DFn	DFd	F	p	part. η^2
ICA_L	2.85	59.82	1.290	.286	.058
ICA_R	2.73	57.29	0.190	.887	.009
IPA_L	4	84	0.854	.495	.039
IPA_R	4	84	2.053	.094	.089

* Greenhouse-Geisser correction is reflected in the ICA values for both eyes due to violation of the assumption of sphericity.

Measurements for the individual task conditions are summarised across all participants in Table A.2 in the Appendix, which also includes measurements of the baseline. Measurements are relatively stable across the baseline and task conditions. For neither indicator an increase related to the difficulty level can be observed. In some cases (ICA_R and IPA_R), the highest observed value was during the first, easiest task. It can also be seen, that for all cases—except $ICA_R \leftrightarrow$ Task 1—the preceding baseline actually resulted in higher indicator values than the following task condition.

Fig. 2 further shows the distribution of the ICA and IPA values for the task conditions, which confirm the relatively stable levels of the indicators across the task conditions, although the IPA shows slightly more variation than the ICA. The IPA also generally shows a higher level for the right than for the left eye, which is not observed as strongly for the ICA.

The results of the ANOVA are summarised in Table 1 and confirm the observations made above. Neither indicator showed significant results between the task difficulty conditions. ICA_R showed the highest p-value of .887. ICA_L was considerably lower at $p = .286$. A similar disparity between the left and right eye was shown for the IPA, however inverted between the eyes: IPA_L showed a higher significance ($p = .495$) than IPA_R ($p = .094$). Effect sizes also varied greatly between the indicators and eyes, ranging from a low effect size for ICA_R, to medium effect sizes for ICA_L, IPA_L, and IPA_R.

3.2. ICA and IPA under the influence of screen luminance

Measurements of both indicators for conditions of bright and dark luminance are listed in Table A.3 in the Appendix, which again includes the baseline measurements. For the ICA the measured values clearly contradict the claims of being robust towards differing light influences.

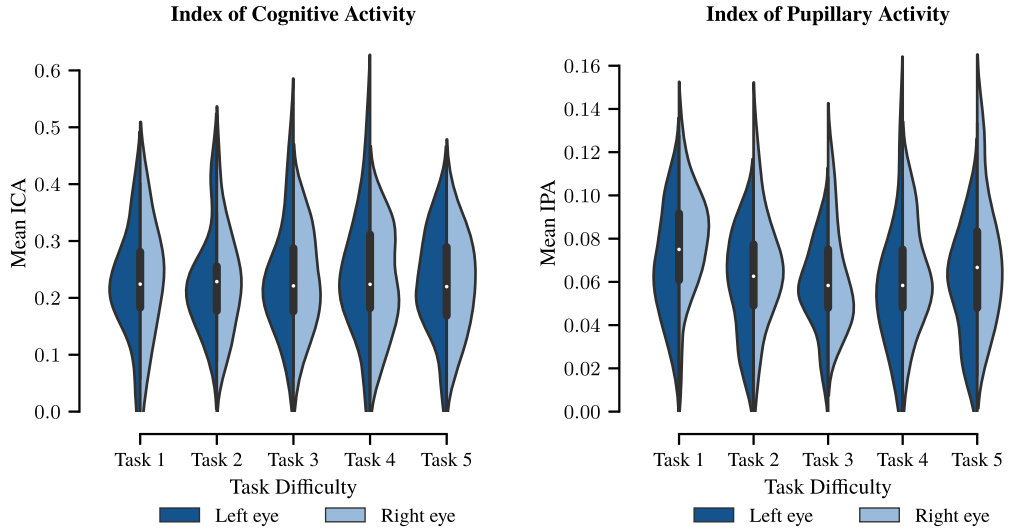


Fig. 2. Violin plots of ICA and IPA values for the task difficulty conditions. The center bar represents the interquartile range, the white dot the median, of both eyes. Distributions are cut off at 0, as the ICA and IPA cannot take negative values.

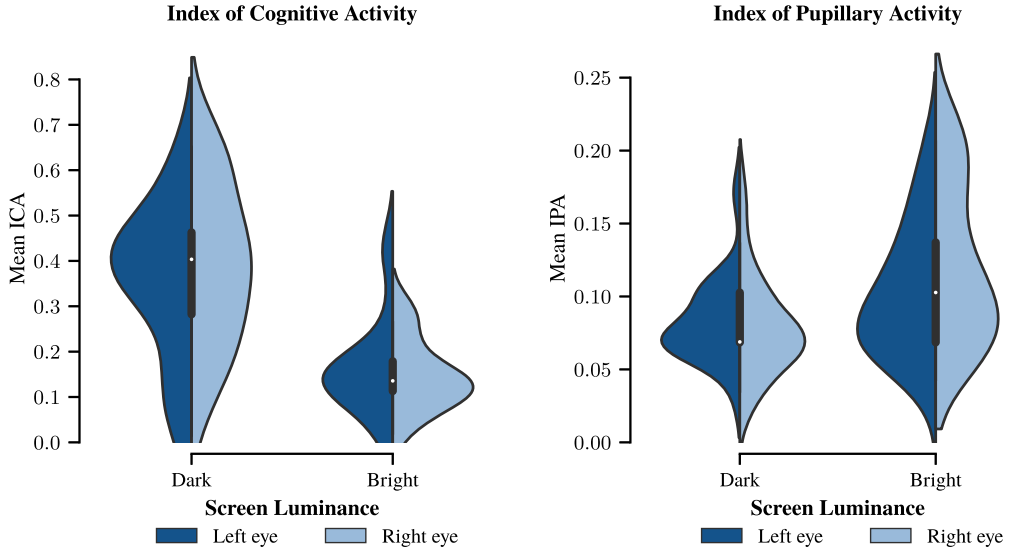


Fig. 3. Violin plots of ICA and IPA values for different screen luminance conditions. The center bar represents the interquartile range, the white dot the median, of both eyes. Distributions are cut off at 0, as the ICA and IPA cannot take negative values.

Mean values for dark luminance are higher than those for the bright luminance condition, for both the left (AM = 0.355 ↔ AM = 0.170) and the right (AM = 0.370 ↔ AM = 0.155) eye. Moreover, the ICA measured during the baseline is higher than both of the following conditions.

While not as large, a difference between the conditions can also be observed for the IPA. The observation is, however, inverse, in that the bright luminance shows a higher value than the dark luminance condition. The baseline measurements are the lowest overall. These observations are also reflected in Fig. 3.

Results of the Wilcoxon-test and effect sizes are summarised in Table 2. The ICA showed highly significant results for both eyes (p = .003 and p = .001 for the left and right eyes, respectively), each with large effect sizes. To the contrary, results for the IPA are mixed between the eyes. For IPA_R the Wilcoxon test showed significant results (p = .007) with a large effect size. IPA_L, however, was not significant (p = .089) and the weakest effect size of all observations. Due to the significance of all but one of the results, additional equivalence tests were not carried out.

Table 2
Wilcoxon Test Statistics.

Indicator	Z	p	Effect Size r
ICA_L	154	.003	.703
ICA_R	160	.001	.765
IPA_L	59	.089	.384
IPA_R	32	.007	.609

4. Discussion

4.1. ICA and IPA as indicators of cognitive strain

According to the results neither indicator was able to measure cognitive strain induced by the mental arithmetic tasks. Participants were presented with five collections of problems with increasing difficulty, which was reflected in the subjective strain assessment as well as the performance measures. The increase in cognitive strain, which is expected from increasing task difficulty, was not differentiated by the indicators. To the contrary, for IPA_L and IPA_R the first condition ac-

tually showed the highest value overall. The same is true for ICA_R. In all but one case (ICA_R:Task/Base1), the baseline measurements taken before each task condition actually showed higher values than the task itself. This means that not only were the indicators not able to differentiate between different levels of cognitive strain but that they were also not able to differentiate between states of no strain versus a situation in which cognitive strain was given.

Assuming a perfect study design and experimental conditions this would indicate that both the ICA and the IPA do not reflect cognitive strain. Based on results from previous work, this would not be the first study to disprove this (e.g., Korbach et al. (2017, 2018); Vogels et al. (2018); Mantzke et al. (2014) for the ICA; Fehringer (2020); Książek et al. (2021) for the IPA). However, certain effects were witnessed which could potentially have led to these results. For one, the order of the tasks was not randomised and, although not explicitly mentioned, participants could have anticipated that each set of tasks would be more challenging than the previous. This could have influenced the given answers for the subjective strain assessment after each condition. Participants could knowingly or unknowingly have adjusted their answers to meet the expectations of the study design and thus proven the trend of increasing difficulty even though they were already overstrained. This is an often observed type of response bias and a phenomenon referred to as *acquiescence*, which describes the tendency of humans to agree to the presented (Paulhus, 1991). However, increasing difficulty was also reflected in the performance measures. The lack of reflection in the indicators could therefore also be due to other emotions or effects reflected in the size of the pupil. Previous studies have shown a connection between pupillary changes and emotional arousal (Bradley et al., 2008; Henderson et al., 2018; Hess and Polt, 1960). One such an emotion could be nervousness or anxiety, which could explain the higher baseline values measured before the task. The staring at a blank white screen while waiting for a task to begin, which is sure to be more difficult than the previous one, can understandably cause a certain amount of emotional stress. This effect could have been further enhanced by the fact that the experiment observer was sat next to the participant. Results for the answers were given orally and although no feedback regarding the correctness of the answer was given, this direct feeling of judgement could have again led to an increase in anxiety based on performance expectations.

Thus, under the assumption that the study design led to an increased baseline level of cognitive strain for the participants even before the conditions started, this could explain why no further increase in strain was detected. If we assume that the indicators themselves deliver on their claims and are capable of detecting user cognitive strain, this has two implications. First, this could mean that either the indicators are not able to detect a further increase in cognitive strain once a certain level has already been reached. That this level of cognitive overload is reached even before the first task condition solely due to the anticipation of the experimental tasks is, however, highly unlikely. Furthermore, in this case, the effectiveness of the indicators are limited regardless, because normal, everyday situations would quickly lead to them being rendered ineffective. The second, more likely implication is that the indicators are only able to detect conditions of cognitive strain versus a condition of no strain. While still useful, this would somewhat hamper the versatility of the indicators. These findings are also indicated in some publications found in previous research, e.g., Fehringer (2020) for the IPA, and could possibly be the causes for the positive results when comparing novice to expert users, e.g., Nguyen et al. (2020); Richstone et al. (2010); Gunawardena et al. (2019).

Another explanation could be that the indicators are sensitive to cognitive strain induced by only a certain type of task. Overall, arithmetic tasks have previously been shown to reflect cognitive strain as measured by the pupil diameter (Beatty, 1982; Marquart and de Winter, 2015; de Winter et al., 2021). However, based on the presented results, the same cannot be necessarily said for the ICA or the IPA. This is also suggested by the meta-analysis of the ICA, in which moderator analyses showed significant effects for the type of cognitive task used (Czerniak et al., 2021a). The basis for comparison is limited, as only a few studies of the ICA have been conducted using arithmetic tasks, e.g., Marshall et al. (2004); Marshall (2007). For the IPA, none could be found.

A phenomenon observed in previous research (e.g., Engonopulos et al. (2013)) was a difference between the ICA values between the left and right eyes. Depending on the task, the indicator was shown to be more or less sensitive. This was not observed in the presented results, which indicate that there is no difference between the left or right eye. This is true for both the ICA and the IPA.

In summary, the findings do not provide evidence regarding the effectiveness of the ICA or the IPA as a robust indicator of cognitive strain. Based on the statistical analysis, neither of the null-hypotheses, regarding the measurement of cognitive strain, are rejected: the manipulation of cognitive tasks of increasing difficulty and thus increasing cognitive strain was not reflected by either indicator at a statistically significant level.

4.2. ICA and IPA under the influence of screen luminance

The results for the effects of the screen luminance on the indicators overall also do not support the claims of the ICA to be robust towards such influences. The screen luminance conditions clearly had an effect on the values of the ICA, which varied greatly depending on whether the screen background is a dark or a bright one. The picture for the IPA is a bit more ambiguous. The distribution of the values is not as clear as for the ICA, which is also reflected in the results of the Wilcoxon-test. Test results for IPA_L were not statistically significant. IPA_R, on the other hand, showed significant test results, indicating a difference between the bright and dark luminance conditions and further showing different behaviour between the left and right eyes. So while the ICA was definitely influenced by the conditions of screen luminance, the IPA remained relatively unaffected.

Interestingly, the values of the ICA seemed to follow the behaviour of the pupils, whereby the values during the dark luminance are higher than those during the bright condition. Assuming that the indicators are in fact able to measure cognitive strain, this would indicate that a higher level of cognitive strain occurred in front of a dark screen, although no task was given. This is relatively unlikely, so that a more plausible explanation would be that the ICA only unsuccessfully removes the pupil dilations caused by luminance effects and thus follows the light reflexes of the pupil diameter.

The IPA on the other hand shows the opposite effect. Significance of the Wilcoxon-test aside, mean values for the bright condition are generally higher than those of the dark one. At the very least this shows a different relation to the pupil diameter than the ICA, which is unexpected. It could also be an indication of a different underlying phenomenon: that the screen luminance itself has an effect on cognitive strain as measured by the IPA. Computer Vision Syndrome (CVS), for example, encompasses a number of different ocular complaints which are associated with the (prolonged) looking at a computer screen (Blehm et al., 2005). These complaints include conditions such as eye strain, dryness, or fatigue, but can also result in non-ocular complaints such as headaches. An effect on cognitive resources is thus plausible. Darker backgrounds and screens (also referred to as “dark mode” or “negative polarity”) have been shown to alleviate such effects (Kim et al., 2019; Erickson et al., 2020). It could be possible that the brighter, white background is in general more strenuous than the darker, black background on the computer screen. Somewhat contradicting these considerations, however, are studies by Piepenbrock et al. (2013, 2014), which showed that during proofreading tasks, dark text on lighter background led to higher performance than if lighter text was presented on a darker background. They refer to this phenomenon as the “positive polarity advantage”. Pedersen et al. (2020), however, found no difference in the performance of a visually intensive screen typing task between a dark and light mode. Both cases, could, nevertheless, leave room for small

influences of the screen luminance on the cognitive strain, and leave an outlook for a phenomenon which should be examined further. Especially when considering influences of differing lighting conditions on the indicators, screen luminance and ambient lighting should, at least in a first step, be treated separately. This applies to both the ICA and the IPA. Previous examinations by the respective authors have shown that both indicators are robust against influences of ambient lighting (Marshall, 2007; Duchowski et al., 2018). The present results, however, indicate that this might not be the case for screen luminance, which could have implications for the overall suitability. Should, for example, the indicators successfully eliminate the effects of ambient lighting, but not those of screen luminance, their use would be limited to use cases which do not involve the prolonged gazing upon a computer screen. This would be unfortunate, as the use of the indicators in such scenarios is highly promising.

In summary, based on the statistical analysis both null-hypotheses to the influences of screen luminance on the indicators are rejected: the manipulation of the screen luminance between dark and bright conditions was reflected in both indicators. While not as inconclusive as the first analysis, the present results do not allow an assessment to whether the IPA can reliably be used as an alternative to the ICA. The ICA is clearly influenced by the examined screen luminance conditions. The IPA, on the other hand, shows more promise in terms of its robustness towards the screen luminance, although the considerations above require further investigation.

5. Limitations

Being a secondary analysis of data collected in previous studies, the analysis faces the same limitations which occurred in the primary study design and execution. This includes the lack of randomisation of the tasks or the potential nervousness of participants in Study I, as discussed above. For Study II, while a randomisation of the conditions was given, other images and perceptional effects were used. The presented analysis only includes two images for the dark and bright luminance conditions and thus only contains a section of the full study design. It cannot be said for certain that the other images did not have an effect on the measurements, even though the first second of the measurements was removed to eliminate direct influences. Furthermore, both studies were carried out with a sample size of 22 participants in Study I and 20 participants for Study II, the data of some of which had to be removed for individual statistical tests. While most comparable studies to the ICA are conducted with similar sample sizes (cf. Czerniak et al. (2021a)) and the conducted sensitivity analyses to the study power show that the sample sizes overall are sufficient to detect at least medium effects, this should be considered especially with regard to the non-significant test results for IPA_L in Study II. Test parameters of the sensitivity analyses can be found in Table A.4 in the Appendix.

Another limitation both studies faced is the general sensitivity of eye tracking data. Even though the studies took precautions to ensure the best possible data collection (such as chin rests to stabilise the participants' gaze in Study II) pupillometric data often still suffers from noise and measurement errors (Kret and Sjak-Shie, 2019). Differences in the way the data is processed can thus have potentially large effects on the results. One example is the detection and handling of the blinks, which is described in the Appendix (p. 8). Other methods for the blink detection, however, also exist, (e.g., Geller et al. (2020); Al-gawwam and Benaissa (2018); Hershman et al. (2018)). Methods for the removal of blinks can also differ to exclude more or less data points before and after the blinks—in some cases removing just the blink itself—or the linear interpolation of the two valid measurements before and after a blink. More technically complex methods include the algorithmic analysis of the noise and smoothing of the pupillary signal (Al-gawwam and Benaissa, 2018). While these methods were not examined as part of this work, they provide possible points for adjustment, especially of the IPA, which in this case benefits from its openness.

6. Conclusions

The aim of this work was to examine the differences between the two pupillometric measures ICA and IPA in the context of cognitive strain measurement (RQ 1) and under the influence of screen luminance (RQ 2). Based on the results, an assessment of whether the IPA offers a viable alternative or replacement to the ICA was to be made. Unfortunately, in none of the cases a definitive and clear-cut answer can be given.

Regarding RQ 1, at least in the presented study, no relevant difference between the ICA and the IPA is observed. Neither indicator was able to differentiate between five levels of cognitive strain as induced by mental arithmetic tasks. The increasing difficulty was shown in the performance measures to the task as well as the subjective strain assessment by the participants. The same trend is not observed for the ICA or the IPA and in some cases baseline measurements exceed those of the task condition. However, it is also possible that certain aspects of the study design led to this result; for example, the lack of randomisation or the presence of the study observer during the conditions. Another possibility is that the indicators reflected the nervousness of the participants.

Results show a similar, yet slightly different picture regarding RQ 2. Overall, the examinations to the robustness towards influences of screen luminance are also in contradiction to the claims. The ICA and IPA values clearly differed between conditions of dark and bright screen luminance. Interestingly, the ICA seems to mirror the behaviour of the pupil, while the IPA showed the opposite effect. It is thus possible that while the ICA fails to eliminate any screen luminance influences, the IPA detects strain which is based on perceptional effects and fatigue associated with brighter computer screens.

In summary, the claims of the pupillary measures ICA and IPA were not confirmed with the data from the used studies. This applies to the ability of the indicators to measure different levels of cognitive strain as well as their robustness towards lighting influences. It is clear that further research is required before either measure can be used as a definitive indicator of cognitive strain. A starting point for further investigations could be the examinations of the indicators in differing conditions of both ambient and screen luminance as well as other types of cognitive tasks. If and when the claims of the indicators can be proven and different levels of cognitive strain are reliably detected, even under varying lighting conditions, the next step should be to examine more complex tasks, which mimic or replicate actual work tasks. Such use case specific examinations will be necessary to develop indicators that provide reliable measurements of cognitive strain in real world settings.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The underlying studies were funded by the European Union's Horizon 2020 Research and Innovation Programme under the grant agreements No 723373 and No 826232, within the *"Smart and Adaptive Interfaces for INCLUSIVE Work Environment"* and *"Smart Working Environments for all Ages (WorkingAge)"* projects, respectively. Additional analyses and transfer of the gained insights to industry practices are carried out within the *"Demonstrations- und Transfernetzwerk KI in der Produktion (ProKI-Netz)"* initiative, funded by the German Federal Ministry of Education and Research (BMBF) under the grant number 02P22A010.

Appendix A

Definition: Cognitive strain

The term *cognitive strain* used in this work is based on a American Psychological Association (2022) definition, whereby *cognition* includes *"all forms of knowing and awareness, such as perceiving, conceiving, remembering, reasoning, judging, imagining, and problem solving."* Cognitive processes thus describe a specific kind of mental processing, which is closer associated with the informational work and phases of non-physical work as defined by Rohmert (1983); Schlick et al. (2018). *Strain* is used based on the distinctions made in Rohmert's (1984) stress and strain concept whereby strain describes the individual demands placed on a person by external stress factors.

Table A.1
Overview of multiplicands used in the mental arithmetic tasks, Study I. Adapted from Czerniak et al. (2021b).

Task Difficulty	1st Multiplicand	×	2nd Multiplicand
Task 1	[6, 9]	×	[6, 9]
	[6, 9]	×	[11, 14]
Task 2	[6, 9]	×	[6, 9]
	[11, 14]	×	[11, 14]
Task 3	[11, 14]	×	[16, 19]
	[16, 19]	×	[16, 19]
Task 4	[21, 24]	×	[26, 29]
	[26, 29]	×	[26, 29]
Task 5	[31, 34]	×	[36, 39]
	[36, 39]	×	[36, 39]

The task difficulty levels are in order of increasing difficulty from Task 1 to Task 5, each with a total of 22 multiplication problems. Squared products were removed from the set of presented problems, i.e., the first and seconds multiplicands are not equal.

Data preparation and calculations

The used eye tracker records a pupil diameter value of zero (0) if no pupil can be detected. This occurs either if the participant looks away from the remotely placed eye tracker or if the eyes are closed (e.g., during a blink). The first action was accounted for during the experiments as participants held their gaze towards the screen for the entire duration of the experimental conditions. Therefore, all recorded zero values can be attributed to either measurement errors or blinks. Using the average duration of a blink—0.1-0.4 seconds (Milo et al., 2010, BNID 100706)—a series of zero values are classified either as blinks or as errors based on their length.

The detection of blinks plays an important role for the preprocessing for the calculation of the IPA. Hereby, 200 ms before and after the identified blinks are marked for removal and deleted for the final signal. This is the approach as described by Duchowski et al. (2018) in the publication to the IPA. The removal of these values is done to cleanse the signal from false pupil diameter measurements occur during the opening and closing of the eyelids (Hershman et al., 2018). Through the partial covering of the pupil by the eyelids the detected area of the pupil can be much smaller than the actual size and thus led to artificial effects.

For Study II an additional step of cleansing the pupil diameter signal was necessary. This is due to the setup of the original experiment, in which images cycled through automatically and conditions succeeded each other without a break in between (Czerniak et al., 2021b). In some cases an effect of the previous image could be observed in the pupil diameter signal, for example if the white background image was shown directly after an image with a black background resulting in a much larger pupil diameter at the beginning of the condition. This artifact skews the average pupil diameter to a higher level than it actually is. Therefore, the first few measurements of the condition must be removed. Based on the fact that the pupillary response to an introduced stimulus can take up to 1 second (Beatty and Lucero-Wagoner, 2000), the pupil diameter signal is shortened by this length.

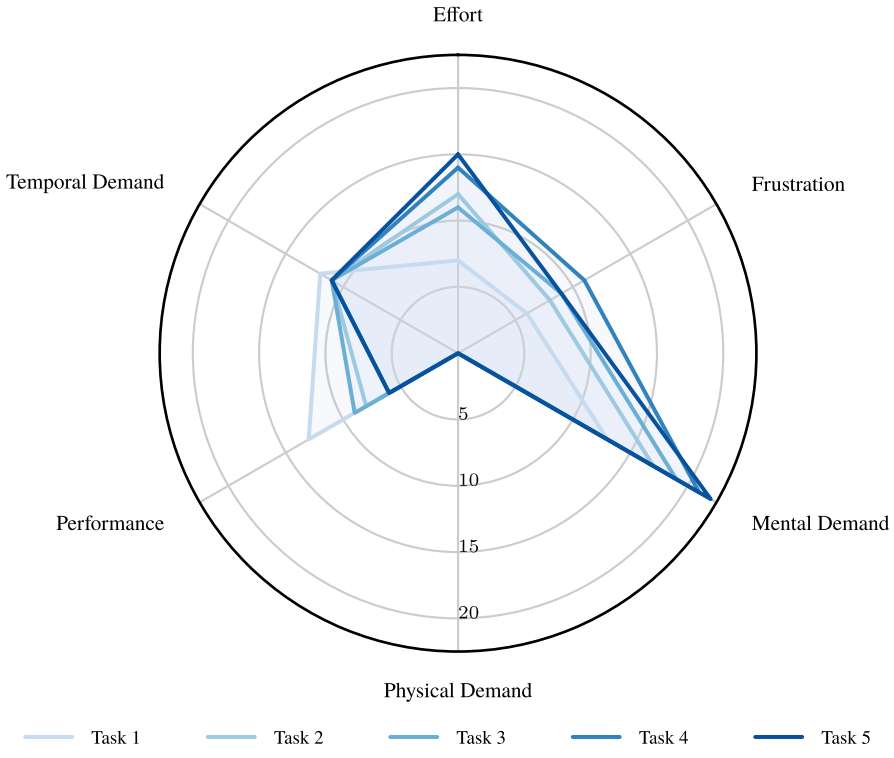


Fig. A.1. Mean weighted scores of the NASA-TLX items for different task difficulty conditions.

Table A.2
Arithmetic mean and standard deviation of ICA (N = 22) and IPA (N = 22) values for different task difficulty conditions.

	ICA_L		ICA_R		IPA_L		IPA_R	
Condition	AM	SD	AM	SD	AM	SD	AM	SD
Baseline								
Base 1	0.234	0.009	0.225	0.108	0.091	0.033	0.099	0.044
Base 2	0.245	0.081	0.240	0.098	0.111	0.049	**0.128**	0.052
Base 3	**0.254**	0.092	**0.265**	0.099	0.111	0.045	0.111	0.043
Base 4	0.245	0.082	0.257	0.105	**0.126**	0.042	0.109	0.037
Base 5	0.249	0.123	0.257	0.123	0.114	0.054	0.108	0.053
Task Difficulty								
Task 1	0.222	0.092	**0.234**	0.101	**0.068**	0.024	**0.081**	0.023
Task 2	0.228	0.087	0.228	0.088	0.059	0.023	0.067	0.025
Task 3	0.237	0.099	0.231	0.086	0.057	0.019	0.063	0.024
Task 4	**0.244**	0.110	0.228	0.089	0.057	0.029	0.064	0.024
Task 5	0.230	0.088	0.227	0.084	0.058	0.024	0.072	0.030

Bold values indicate the highest value for the respective indicator and eye.

Table A.3
Arithmetic mean and standard deviation of ICA (N = 18) and IPA (N = 20) values for different screen luminance conditions.

	ICA_L		ICA_R		IPA_L		IPA_R	
Condition	AM	SD	AM	SD	AM	SD	AM	SD
Baseline	**0.430**	0.165	**0.437**	0.144	0.071	0.008	0.071	0.008
Dark Luminance	0.355	0.167	0.370	0.192	0.083	0.028	0.079	0.033
Bright Luminance	0.170	0.111	0.155	0.076	**0.103**	0.044	**0.122**	0.054

Bold values indicate the highest value for the respective indicator and eye.

Table A.4
Results of the sensitivity analyses to the study power.

Study / Indicator	Sample Size	Alpha	Power	Effect Size*	Classification**
Study I: ANOVA Test					
ICA & IPA	22	0.05	0.95	0.298	medium effect
			0.80	0.240	small effect
Study II: Wilcoxon Test					
ICA	18	0.05	0.95	0.926	large effect
			0.80	0.719	medium effect
IPA	20	0.05	0.95	0.873	large effect
			0.80	0.678	medium effect

* Effect size f for ANOVA, d for Wilcoxon Tests.
** Based on classification in Cohen (1988).

The subsequent calculation of the IPA was done using the python scripts provided by Duchowski et al. (2018). The only change to the original scripts was the substitution of the symlet 16 wavelet through a symlet 4 wavelet to account for the lower measurement frequency of 60 Hz of the FOVIO FX3 eye tracker. Calculation of the ICA was carried out using the required, proprietary EyeWorks Cognitive Workload module (EyeTracking, 2020).

Other than the methods described, no further removal or cleansing of the data was done. All IDs which showed a valid pupil diameter signal were processed and respected in the statistical analysis. All data preparation and processing was done in python using the numpy and pandas packages (Reback et al., 2021; Harris et al., 2020). The statistical analyses were carried out in R using the tidyverse and rstatix libraries (Wickham et al., 2019; Kassambara, 2021). Power analyses were carried out using G*Power (Faul et al., 2007).

References

Ahern, S., Beatty, J., 1979. Pupillary responses during information processing vary with Scholastic Aptitude Test scores. Science 205 (4412), 1289–1292. https://doi.org/10.1126/science.472746.

Al-gawwam, S., Benaissa, M., 2018. Robust eye blink detection based on eye landmarks and Savitzky–Golay filtering. Information 9 (4), 93. https://doi.org/10.3390/info9040093.

American Psychological Association, 2022. Definition of cognition. APA Dictionary of Psychology. Retrieved March 2, 2022, from http://dictionary.apa.org/cognition.

Armougum, A., Gaston-Bellegarde, A., Joie-La Marle, C., Piolino, P., 2020. Physiological investigation of cognitive load in real-life train travelers during information processing. Appl. Ergon. 89. https://doi.org/10.1016/j.apergo.2020.103180.

Bartels, M., Marshall, S.P., 2012. Measuring cognitive workload across different eye tracking hardware platforms. In: Morimoto, C.H. (Ed.), Proceedings of the Symposium on Eye Tracking Research and Applications. ACM, pp. 161–164.

Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. 91 (2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276.

Beatty, J., Lucero-Wagoner, B., 2000. The pupillary system. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G. (Eds.), Handbook of Psychophysiology. Cambridge University Press, pp. 142–162. http://psycnet.apa.org/record/2000-03927-005.

Blehm, C., Vishnu, S., Khattak, A., Mitra, S., Yee, R.W., 2005. Computer vision syndrome: a review. Survey Ophthalmol. 50 (3), 253–262. https://doi.org/10.1016/j.survophthal.2005.02.008.

Bortz, J., Schuster, C., 2010. Statistik für Human- und Sozialwissenschaftler, 7th ed. Springer.

Bradley, M.M., Miccoli, L., Escrig, M.A., Lang, P.J., 2008. The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology 45 (4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x.

Charles, R.L., Nixon, J., 2019. Measuring mental workload using physiological measures: a systematic review. Appl. Ergon. 74, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. L. Erlbaum Associates.

Czerniak, J.N., Schierhorst, N., Brandl, C., Mertens, A., Schwalm, M., Nitsch, V., 2021a. A meta-analytic review of the reliability of the Index of Cognitive Activity concerning task-evoked cognitive workload and light influences. Acta Psychol. 220, 103402. https://doi.org/10.1016/j.actpsy.2021.103402.

Czerniak, J.N., Schierhorst, N., Villani, V., Sabattini, L., Brandl, C., Mertens, A., Schwalm, M., Nitsch, V., 2021b. The Index of Cognitive Activity - Eligibility for task-evoked informational strain and robustness towards visual influences. Appl. Ergon. 92, 103342. https://doi.org/10.1016/j.apergo.2020.103342.

Debue, N., van de Leemput, C., 2014. What does germane load mean? An empirical contribution to the cognitive load theory. Front. Psychol. 5. https://doi.org/10.3389/fpsyg.2014.01099.

Devos, H., Gustafson, K., Ahmadnezhad, P., Liao, K., Mahnken, J.D., Brooks, W.M., Burns, J.M., 2020. Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults. Brain Sci. 10 (12), 994. https://doi.org/10.3390/brainsci10120994.

de Winter, J.C.F., Petermeijer, S.M., Kooijman, L., Dodou, D., 2021. Replicating five pupillometry studies of Eckhard Hess. Int. J. Psychophysiol. https://doi.org/10.1016/j.ijpsycho.2021.03.003.

Duchowski, A.T., 2017. Eye Tracking Methodology: Theory and Practice. Springer International Publishing.

Duchowski, A.T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., Giannopoulos, I., 2018. The index of pupillary activity: measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–13.

Engonopulos, N., Sayeed, A., Demberg, V., 2013. Language and cognitive load in a dual task environment. Proc. Ann. Meeting Cogn. Sci. Soc. 35.

Erickson, A., Kim, K., Bruder, G., Welch, G.F., 2020. Effects of dark mode graphics on visual acuity and fatigue with virtual reality head-mounted displays. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 434–442.

EyeTracking, Inc., 2014. Workload Module Manual. (Version 1.0).

EyeTracking, Inc, 2020. EyeWorks software - cognitive workload modules. Retrieved April 26, 2022, from http://www.eyetracking.com/cognitive-workload/.

EyeTracking, Inc., 2021. FX3 - the versatile turnkey eye tracking solution. EyeTracking. Retrieved April 26, 2022, from http://www.eyetracking.com/fx3-remote-eye-tracking/.

EyeTracking, Inc., 2022. EyeWorks software - advanced and full featured eye tracking solution. EyeTracking. Retrieved April 26, 2022, from http://www.eyetracking.com/eyeworks-software/.

Fairclough, S.H., Moores, L.J., Ewing, K.C., Roberts, J., 2009. Measuring task engagement as an input to physiological computing. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–9.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191. https://doi.org/10.3758/BF03193146.

Fehringer, B.C.O.F., 2020. One threshold to rule them all? Modification of the Index of Pupillary Activity to optimize the indication of cognitive load. In: ACM Symposium on Eye Tracking Research and Applications, pp. 1–5.

Geller, J., Winn, M.B., Mahr, T., Mirman, D., 2020. GazeR: a package for processing gaze position and pupil size data. Behav. Res. 52 (5), 2232–2255. https://doi.org/10.3758/s13428-020-01374-8.

Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. Psychometrika 24 (2), 95–112. https://doi.org/10.1007/BF02289823.

Gunawardena, N., Matscheko, M., Anzengruber, B., Ferscha, A., Schobesberger, M., Shamiyeh, A., Klugsberger, B., Solleder, P., 2019. Assessing surgeons' skill level in laparoscopic cholecystectomy using eye metrics. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585 (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Advances in Psychology. North-Holland, pp. 139–183.

Henderson, R.R., Bradley, M.M., Lang, P.J., 2018. Emotional imagery and pupil diameter. Psychophysiology 55 (6), e13050. https://doi.org/10.1111/psyp.13050.

Hershman, R., Henik, A., Cohen, N., 2018. A novel blink detection method based on pupillometry noise. Behav. Res. 50 (1), 107–114. https://doi.org/10.3758/s13428-017-1008-1.

Hess, E.H., Polt, J.M., 1960. Pupil size as related to interest value of visual stimuli. Science 132 (3423), 349–350. https://doi.org/10.1126/science.132.3423.349.

Hess, E.H., Polt, J.M., 1964. Pupil size in relation to mental activity during simple problem-solving. Science 143 (3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190.

Holmqvist, K., 2017. Eye Tracking: A Comprehensive Guide to Methods, Paradigms, and Measures.

Kahneman, D., Beatty, J., 1966. Pupil diameter and load on memory. Science 154 (3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583.

Kahya, M., Wood, T.A., Sosnoff, J.J., Devos, H., 2018. Increased postural demand is associated with greater cognitive workload in healthy young adults: a pupillometry study. Front. Human Neurosci. 12, 288. https://doi.org/10.3389/fnhum.2018.00288.

Kassambara, A., 2021. Package 'rstatix': pipe-friendly framework for basic statistical tests (CRAN package). http://rpkgs.datanovia.com/rstatix/.

Kim, K., Erickson, A., Lambert, A., Bruder, G., Welch, G., 2019. Effects of dark mode on visual fatigue and acuity in optical see-through head-mounted displays. In: Symposium on Spatial User Interaction, pp. 1–9.

Korbach, A., Brünken, R., Park, B., 2017. Measurement of cognitive load in multimedia learning: a comparison of different objective measures. Instr. Sci. 45 (4), 515–536. https://doi.org/10.1007/s11251-017-9413-5.

Korbach, A., Brünken, R., Park, B., 2018. Differentiating different types of cognitive load: a comparison of different measures. Educ. Psychol. Rev. 30 (2), 503–529. https://doi.org/10.1007/s10648-017-9404-8.

Kret, M.E., Sjak-Shie, E.E., 2019. Preprocessing pupil size data: guidelines and code. Behav. Res. 51 (3), 1336–1342. https://doi.org/10.3758/s13428-018-1075-y.

Książek, P., Zekveld, A.A., Wendt, D., Fiedler, L., Lunner, T., Kramer, S.E., 2021. Effect of speech-to-noise ratio and luminance on a range of current and potential pupil response measures to assess listening effort. Trends Hear 25, 23312165211009351. https://doi.org/10.1177/23312165211009351.

Laeng, B., Alnaes, D., 2019. Pupillometry. In: Klein, C., Ettinger, U. (Eds.), Eye Movement Research: An Introduction to its Scientific Foundations and Applications, 1st ed. Springer, pp. 449–502.

Lakens, D., 2017. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc. Psychol. Pers. Sci. 8 (4), 355–362. https://doi.org/10.1177/1948550617697177.

Mantzke, O., Vilimek, R., Carbon, C.C., 2014. Driver distraction: Weiterentwicklung einer methode zur messung von kognitiver beanspruchung durch anzeige-und bedienkonzepte im fahrzeug. Grundlagen Und Anwendungen Der Mensch-Maschine-Interaktion, vol. 10. Berliner Werkstatt Mensch-Maschine-Systeme.

Marquart, G., Cabrall, C., de Winter, J., 2015. Review of eye-related measures of drivers' mental workload. Proc. Manufact. 3, 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783.

Marquart, G., de Winter, J., 2015. Workload assessment for mental arithmetic tasks using the task-evoked pupillary response. PeerJ Comput. Sci. 1, e16. https://doi.org/10.7717/peerj-cs.16.

Marshall, S.P., 2000. Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. (U.S. pat. 006090051A). http://patents.google.com/patent/US6090051A/en.

Marshall, S.P., 2007. Identifying cognitive state from eye metrics. Aviat. Space Environ. Med. 78 (5), B165–B175. http://www.ingentaconnect.com/content/asma/asem/2007/00000078/a00105s1/art00025.

Marshall, S.P., Davis, C.L., Knust, S.R., 2004. The Index of Cognitive Activity: Estimating Cognitive Effort from Pupil Dilation. Technical Report. EyeTracking, Inc., San Diego, CA, USA.

Mauchly, J.W., 1940. Significance test for sphericity of a normal n-variate distribution. Ann. Math. Stat. 11 (2), 204–209. https://doi.org/10.1214/aoms/1177731915.

Milo, R., Jorgensen, P., Moran, U., Weber, G., Springer, M., 2010. BioNumbers—the database of key numbers in molecular and cell biology. Nucleic Acids Res. 38, D750–D753. https://doi.org/10.1093/nar/gkp889.

Nguyen, J.H., Chen, J., Marshall, S.P., Ghodoussipour, S., Chen, A., Gill, I.S., Hung, A.J., 2020. Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. World J. Urol. 38 (7), 1599–1605. https://doi.org/10.1007/s00345-019-02881-w.

Paulhus, D.L., 1991. Measurement and Control of Response Bias. Measures of Personality and Social Psychological Attitudes. Elsevier, pp. 17–59.

Pedersen, L.A., Einarsson, S.S., Rikheim, F.A., Sandnes, F.E., 2020. User interfaces in dark mode during daytime – improved productivity or just cool-looking? In: Antona, M., Stephanidis, C. (Eds.), Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies. Springer International Publishing, pp. 178–187.

Peysakhovich, V., Causse, M., Scannella, S., Dehais, F., 2015. Frequency analysis of a task-evoked pupillary response: luminance-independent measure of mental effort. Int. J. Psychophysiol. 97 (1), 30–37. https://doi.org/10.1016/j.ijpsycho.2015.04.019.

Piepenbrock, C., Mayr, S., Buchner, A., 2014. Smaller pupil size and better proofreading performance with positive than with negative polarity displays. Ergonomics 57 (11), 1670–1677. https://doi.org/10.1080/00140139.2014.948496.

Piepenbrock, C., Mayr, S., Mund, I., Buchner, A., 2013. Positive display polarity is advantageous for both younger and older adults. Ergonomics 56 (7), 1116–1124. https://doi.org/10.1080/00140139.2013.790485.

Reback, J., Jbrockmendel, McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Hawkins, S., Gfyoung, Roeschke, M., Sinhrks, Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Hoefler, P., Naveh, S., Garcia, M., Schendel, J., Hayden, A., Saxton, D., Darbyshire, J.H.M., Shadrach, R., Gorelli, M.E., Li, F., Jancauskas, V., McMaster, A., Zeitlin, M., Battiston, P., Seabold, S., 2021. Pandas-dev/pandas: Pandas 1.3.3 (version v1.3.3). https://doi.org/10.5281/ZENODO.3509134.

Rerhaye, L., Blaser, T., Alexander, T., 2018. Evaluation of the Index of Cognitive Activity (ICA) as an instrument to measure cognitive workload under differing light conditions. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). Springer International Publishing, pp. 350–359.

Richstone, L., Schwartz, M.J., Seideman, C., Cadeddu, J., Marshall, S., Kavoussi, L.R., 2010. Eye metrics as an objective assessment of surgical skill. Ann. Surg. 252 (1), 177–182. https://doi.org/10.1097/SLA.0b013e3181e464fb.

Rohmert, W., 1983. Formen menschlicher Arbeit. In: Luczak, H., Rohmert, W., Rutenfranz, J. (Eds.), Was ist Arbeitsphysiologie, pp. 5–29. Retrieved January 30, 2021, from http://tubiblio.ulb.tu-darmstadt.de/45114/.

Rohmert, W., 1984. Das Belastungs-Beanspruchungs-Konzept. Z. Arb.wiss. 38 (4), 193–200. http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=9119422.

Schlick, C., Bruder, R., Luczak, H., 2018. Arbeitswissenschaft. Springer.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

van der Wel, P., van Steenbergen, H., 2018. Pupil dilation as an index of effort in cognitive control tasks: a review. Psychon. Bull. Rev. 25 (6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y.

Vogels, J., Demberg, V., Kray, J., 2018. The index of cognitive activity as a measure of cognitive processing load in dual task settings. Front. Psychol. 9, 2276. https://doi.org/10.3389/fpsyg.2018.02276.

Walker, E., Nowacki, A.S., 2011. Understanding equivalence and noninferiority testing. J. Gen. Intern. Med. 26 (2), 192–196. https://doi.org/10.1007/s11606-010-1513-8.

Weber, P., Rupprecht, F., Wiesen, S., Hamann, B., Ebert, A., 2021. Assessing cognitive load via pupillometry. In: Arabnia, H.R., Ferens, K., de la Fuente, D., Kozerenko, E.B., Olivas Varela, J.A., Tinetti, F.G. (Eds.), Advances in Artificial Intelligence and Applied Cognitive Computing. Springer International Publishing, pp. 1087–1096. http://link.springer.com/10.1007/978-3-030-70296-0_86.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Sour. Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1 (6), 80. https://doi.org/10.2307/3001968.

Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A., 2015. State of science: mental workload in ergonomics. Ergonomics 58 (1), 1–17. https://doi.org/10.1080/00140139.2014.956151.