One threshold to rule them all? Modification of the Index of Pupillary Activity to optimize the indication of cognitive load

Benedict C. O. F. Fehringer
Department of Educational Psychology
University of Mannheim
b.fehringer@uni-mannheim.de

ABSTRACT

Cognitive load is an important source of information in performance situations. One promising non-invasive method is pupillometry. The Index of Pupillary Activity [IPA, Duchowski et al. 2018] performs a wavelet transformation on changes of pupillary dilations to detect high frequencies. This index is inspired by the Index of Cognitive Activity [ICA, Marshall 2000]. The IPA value is the sum of peaks exceeding a predefined threshold. The present study shows that it appears reasonable to adapt this threshold corresponding to the task. Fifty-five participants performed a spatial thinking test with six difficulty levels and two simple fixation tasks. Six different IPA values resulting from different thresholds were computed. The distributions of these IPA values of the eight conditions were analyzed regarding the validity to indicate different levels of cognitive load, corresponding to accuracy data. The analyses revealed that different thresholds are sensitive for different cognitive load levels. Contra-intuitive results were also obtained.

CCS CONCEPTS

• **Human-centered computing** → **Interaction devices**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Pupillometry, Index of Pupillary Activity, cognitive load

ACM Reference Format:
Benedict C. O. F. Fehringer. 2020. One threshold to rule them all? Modification of the Index of Pupillary Activity to optimize the indication of cognitive load. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Short Papers), June 2–5, 2020, Stuttgart, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3379156.3391341

1 INTRODUCTION

Cognitive load is an important information during task performing. It can be used to determine the cognitive demand of certain tasks specific for the current performer. Such information might be utilized to adapt the presented features in automotive systems [Palinko and Kun 2012] or to monitor the cognitive demand during surgery [Zheng et al. 2015].

There are different ways to measure cognitive load, such as blink rate [Chen and Epps 2014], fixation length, heart rate, and electrodermal activity [for an overview see Cowley et al. 2015]. A further promising source of information to measure cognitive load are changes of pupil dilation, which is one of the most valid measures of the non-invasive techniques. However, changes of the pupil's size can either be measured as absolute change (i.e, the wider the pupil is, the higher is the cognitive load), but also considering the pupil's fluctuation (i.e., higher frequencies in pupil changes indicate higher cognitive load), see Andreassi [2007]. One indicator for cognitive load that exploits the last property is the *Index of Cognitive Activity* [ICA, Marshall 2000]. The ICA is a patented algorithm that performs a wavelet transformation on the changes of the pupil dilation. The high frequencies are interpreted to indicate cognitive load. Although there are studies supporting the validity of the ICA [Demberg et al. 2013; Marshall 2009; Schwalm et al. 2008], the ICA is less applied in scientific research supposably due to the patented, and therefore at least partly unpublished algorithm.

An alternative that is inspired by the ICA is the Index of Pupillary Activity [IPA, Duchowski et al. 2018]. The IPA's algorithm is fully open and well documented. So, the algorithm can be optimized with respect to the context in which the cognitive load should be measured. Such a context-dependency might be contra-intuitive, if the algorithm is thought to be rationally derived from neurological and physiological activities. However, it has to be emphasized that the connection from an external input task over basic perception and higher-order mental processes to the innervation of the pupil is not strongly determined and influenced by many other factors, such as fatigue [Andreassi 2007], light [Steinhauer et al. 2004] and individual differences [Zekveld and Kramer 2014]. Just and Carpenter [1993] pointed out that pupillary responses are not causally related with cognitive demand but only correlated. In this sense Kahneman stated that the "dilation of the pupil is the best single *index* [for effort]." [Kahneman 1973, p.18] If a pupillary-based indicator is interpreted as only correlative, but not causally related to cognitive load, then the indicator's algorithm should be optimized with regard to increasing its ability to indicate cognitive load. Hence the algorithm should be validated with external criteria that are known to be related to cognitive demand, such as task difficulty.

In the present study, this optimization was conducted utilizing a validated performance test for spatial thinking, a subdomain of the general intelligence [e.g., Paivio 2014]. The so-called R-Cube-Vis Test [Fehringer 2019] consists of six well-defined difficulty levels being all on the same ability dimension. The optimization procedure for the IPA algorithm was conducted by changing the threshold that determines the highest peaks of the highest frequencies as indication of cognitive load [Duchowski et al. 2018]. This threshold

ETRA '20 Short Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7134-6/20/06…$15.00
https://doi.org/10.1145/3379156.3391341

is fixed for each input data sample in the original version of the IPA. The study goal is to find an optimized threshold with respect to the indication of cognitive load. Hence, different thresholds were considered and evaluated for the six difficulty levels of the R-Cube-Vis Test and a simple fixation task requiring no mental effort.

2 THE INDEX OF PUPILLARY ACTIVITY

The Index of Pupillary Activity [IPA, Duchowski et al. 2018] takes the pupil diameter signal as input and applies a wavelet transformation to decompose the signal. Duchowski et al. [2018] used the symlet-16 wavelets in their analysis and applied them to their eye tracking data from an eye tracker with a sampling rate of 1000Hz. In their following analyses, the detail coefficients of the second level were used. They can be roughly interpreted as the high frequency changes of the pupil diameter with noise removed. In the next step, the algorithm computes the so-called modulus maxima, the modes of the absolute values of the resulting detail coefficients using a sliding window of size 3. If the second sample of this window is greater or equal than the first and third sample and truly greater than at least of one of the first and third sample, the second sample is maintained, otherwise it is set to 0. The resulting modes are compared with a predefined threshold. The IPA value is the number of modes greater than this threshold relative to the considered time interval measured in seconds (= occurrences per second). The threshold used in the algorithm is the universal threshold, λ_{univ}:

$$\lambda_{univ} = \hat{\sigma}\sqrt{(2log_n)} \tag{1}$$

The estimated standard deviation ($\hat{\sigma}$) is based on the modulus maxima of the considered n detail coefficients.

2.1 Study goal

As described above, the threshold utilized to compute the IPA is fixed for each input signal in the original IPA version. However, different thresholds might be more appropriate to indicate cognitive load. As pointed out above, the pupillary reaction should not be understood as a logical derivation of brain activity that would be unambiguously connected to cognitive load. Pupillary reactions should be more seen as a correlate of cognitive load (not causally related) that can be optimized with regard to the interested criterion. Therefore, different thresholds were tested in the current study with respect to their ability to indicate cognitive workload during performing items of different difficulty levels of a performance test (the R-Cube-Vis Test, Section 3.2). This test is well validated with respect to the measured construct and its one-dimensional structure, as the ordered difficulty levels of the R-Cube-Vis Test decrease in their accuracy from the easiest to the most difficult level [Fehringer 2019]. Therefore, this test is suitable as criterion test.

The threshold of the IPA algorithm was varied by multiplying λ_{univ} with the following factors 0; 0.5; 0.8; 1; 1.5; 2. Hence, the final IPA value was the number of the modes greater than the product of λ_{univ} and one of the six factors. That means, that first, factor 1 results in the original algorithm and second, the higher the factor is, the less modes are taken into account for the resulting IPA value. In the present study the symlet-5 wavelets were applied due to the lower sampling rate of the used eye tracker (300Hz). The IPA values were always computed for 1-second windows and then averaged within each task.

It was expected that more difficult tasks demand more cognitive workload. Appropriate threshold settings should be able to mirror this pattern by higher IPA values for more difficult tasks. In this sense, the lowest IPA values should occur during a control task with presumably no demanded cognitive load.

3 METHOD

3.1 Participants

The study was conducted with N = 55 (43 female, 12 male) participants. All participants were students from a German University and received course credit for their participation. Their mean age was M = 21.07 years (SD = 3.84) and ranged from 18 to 39 years.

3.2 Materials

The short version of the R-Cube-Vis test was conducted as performance test. The R-Cube-Vis test was validated as long and short version in four studies described in a dissertation [Fehringer 2019] and measures the visualization ability, which is the main factor of the construct of spatial thinking [Carroll 1993]. Each item of the R-Cube-Vis test shows two Rubik's cubes, a solved one on the left side and a twisted one on the right side (Figure 1). The task is always to decide whether both cubes are equal (possible items) or not equal (impossible items) except of the rotation of single elements.

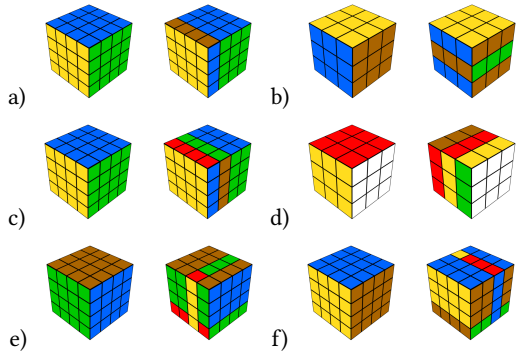

Figure 1: Sample tasks of the R-Cube-Vis Test of the easiest (a) to the most difficult level (f). In all tasks, the right cube could be possibly turned into the left cube.

The R-Cube-Vis Test consists of six distinct difficulty levels, that differ in the size of the cube (4x4x4 vs. 3x3x3 elements), the number of rotated elements (1 vs. 2) and how the two elements are rotated (in parallel vs. crossed). The validation studies demonstrated the existence of the six difficulty levels on a single dimension by conformity of the R-Cube-Vis test with the linear-logistic test model (LLTM). The conducted short version contains 60 items, five possible and five impossible of each difficulty level. The simplest items show cubes of size 4x4x4 with one rotated element (Figure 1a, b). The most difficult items show cubes of size 4x4x4 with two crossed rotated elements (Figure 1e, f). The previous studies with the R-Cube-Vis test showed ceiling effects in the student population for

the simplest items and floor effects for the most difficult items. The test score is the mean of all possible items (0 points for an incorrect answer and 1 point for a correct answer).

Additional to the performance test, a simple fixation test with presumably low cognitive load was conducted before and after the R-Cube-Vis Test. Participants had to fixate on nine crosses presented on an invisible quadratic grid structure consisting on three rows and three columns. The crosses were presented after each other beginning with the cross on the top left and ending with the cross on the bottom right, row by row. Each cross was announced by a square at the same location changing its color from red over yellow to green. Each cross was presented for three seconds.

3.3 Apparatus

The eye tracking data were recorded by a Tobii TX300 eye tracker with 300 Hz recording rate. The eye tracker was integrated in a unit with a screen (resolution of 1920 x 1080 pixels). The distance between participants' eyes and the eye tracker was between 50 and 70cm during recording.

3.4 Procedure

After arriving, the participants were provided an instruction of the experiment followed by calibrating the eye tracker by a 9-point calibration procedure. Then, they performed the first fixation task (*Fixation Task, before*), followed by the R-Cube-Vis Test and the second fixation task (*Fixation Task, after*). At the end, the participants filled out a questionnaire about their sex, age, and study major. The duration of the experiment was between 20 and 35 minutes.

3.5 Analytical approach

For each distribution of the six tested threshold factors, each difficulty level of the R-Cube-Vis was compared with the next difficulty level, i.e., Level a with Level b, Level b with Level c, ..., Level e with Level f. Both fixation tasks were only compared with the easiest difficulty level, Level a, as well as with each other. Each comparison was tested with a directed Bayesian t-test. In case of evidence for equality, an undirected Bayesian t-test was conducted to exploratively test for evidence of equality or of difference in the unexpected direction [see Schönbrodt et al. 2017]. Bayes Factors (BFs) larger than 1 support the difference hypothesis (H_1) and were interpreted as anecdotal ($1 < BF \leq 3$), moderate ($3 < BF \leq 10$), strong ($10 < BF \leq 30$), very strong ($30 < BF \leq 100$) and extreme ($100 < BF$) evident. BFs smaller than 1 support the equality hypothesis (H_0) and were interpreted as anecdotal ($1 > BF \geq 1/3$), moderate ($1/3 > BF \geq 1/10$), strong ($1/10 > BF \geq 1/30$), very strong ($1/30 > BF \geq 1/100$) and extreme ($1/100 > BF$). The resulting distributions of the IPA values were also compared with the distribution of the ICA for comparison. The ICA was computed using the Workload RT software package [Version 1.5, EyeTracking Inc. 2015].

4 RESULTS

The descriptive results confirmed the expected decrease of accuracy and increase of reaction times from the easiest to the most difficult level (Table 1). However, Level a and b and Level c and d were pairwise comparable to each other.

Table 1: Mean and standard deviation of accuracy and reaction times for each difficulty level (see Figure 1)

	Accuracy		Reaction times	
Levels	**M**	**SD**	**M**	**SD**
Level a	.95	.16	5170	2424
Level b	.93	.15	5188	2625
Level c	.86	.20	9871	4806
Level d	.80	.22	10545	5393
Level e	.73	.25	11724	5889
Level f	.42	.30	15714	10707

The following figures (Figure 2 to 4) show the distributions of the referred indicators (IPA or ICA) with their specific parameter setting depending on the performed task. It was expected that the indicator should increase with increasing cognitive load. In contrast, higher indicator values for less demanding tasks would be contra-intuitive. Each distribution shows how the indicator changes if the cognitive load presumably changes. The expected increase of the IPA values from the fixation tasks over the easiest difficulty levels to the most difficulty levels were tested with directed Bayesian T-tests. The original algorithm (threshold factor 1) showed only extreme evidence (BF > 100) between both fixation tasks and Level a from the R-Cube-Vis Test (see Figure 2). The undirected tests showed anecdotal to moderate evidence for equality for all other comparisons except for comparison between Level b and c. However, the extreme evidence for inequality between those levels was in the unexpected direction with higher IPA values for the easier level.

The both smallest threshold factors (0 and 0.5) showed comparable distributions with unexpected high IPA values for both fixation tasks compared to Level a with extreme evidence for inequality (see Figure 2). However the directed tests between Level b and c revealed the expected differences with higher IPA values for the more difficult level with very strong to extreme evidence (BF > 30). For factor 0.5 there was also an unexpected moderate evidence for inequality between Level e and f with lower IPA values for the more difficult level. All other comparison of the factor 0 and 0.5 showed anecdotal to moderate evidence for equality.

Of all comparisons of the two largest factors (1.5 and 2), only the difference between Level e and f showed the expected result with moderate evidence (BF > 3), higher IPA values for the more difficult level. All other comparisons showed either anecdotal evidence or moderate evidence for equality (see Figure 2).

Factor 0.8 produced a distribution with an increasing or stable IPA value from the fixation tasks over the easiest level to the most difficult level (see Figure 2). The comparisons between both fixations tasks and Level a showed strong to very strong evidence for the expected inequality (BF > 10). All other comparisons were either anecdotal or moderate evidence for equality.

There were no differences between both fixation tasks (before and after). For all threshold factors, there were anecdotal to moderate evidence for equality.

The results were more emphasized if the IPA values were z-standardized (see Figure 3). This can be seen by decreased standard errors and more evidence of the conducted Bayesian t-tests.

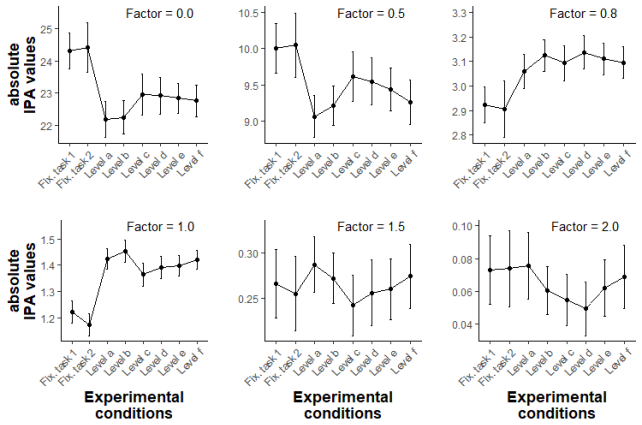

Figure 2: Distribution of the IPA values over the two fixations tasks and the six difficulty levels for the six threshold factors.

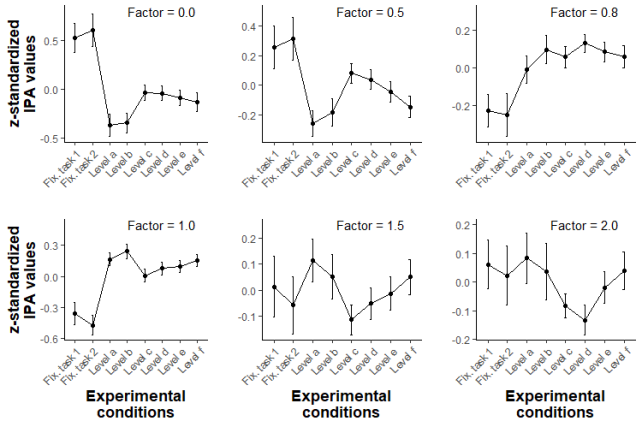

Figure 3: The same distributions as in Figure 2, but with z-standardized IPA-values.

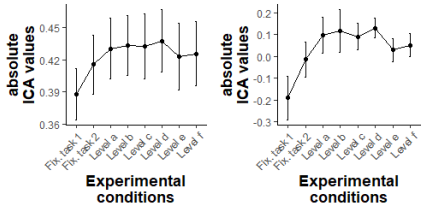

Figure 4: Distribution of the ICA values: absolute and z-standardized.

The distribution of the ICA showed only increasing values from the fixation task before the test to the first difficulty level of the R-Cube-Vis Test (Level a) with extreme evidence. All other comparisons were either anecdotal or evidence for equality, for the absolute as well as for the z-standardized values (Figure 4). Only the comparison between Level d and e showed moderate evidence for inequality with higher values for Level d. The closest similarity of the ICA distributions seems to be with the IPA values using threshold 0.8. Remarkably, the ICA distribution showed also a moderate evidence for inequality between the fixation task *before* the test and the fixation task *after* the test. This might be suggesting a dependence of the ICA from fatigue which could not be found for the IPA.

5 CONCLUSIONS AND FUTURE WORK

The analyses showed the potential of the original version of the IPA (threshold factor 1) to differentiate between non-cognitive-demanding tasks (the fixation tasks) and tasks of the R-Cube-Vis Test as a performance test. However, the different difficulty levels of the R-Cube-Vis Test could not be differentiated by the IPA. Actually, there was a decrease of the IPA between one of the easiest levels (Level b) and one of the medium levels (Level c). However, exactly this comparison was displayed correctly by increasing IPA values for the lowest threshold factors, 0 and 0.5. The expected increase of IPA values between more difficult levels could only be found for the threshold factor 2.0 between the two most difficult levels (Level e and f). Contra-intuitive results were the high IPA values of the fixation tasks for the two lowest threshold factors. A low threshold means that small modes go into the computation of the IPA. In contrast to the cognitively demanding task of spatial thinking, the simple fixation task appears to be characterized by many small modes (see $Factor = 0.0$, Figure 2). The higher threshold has the effect to detect only the large modes. The cognitively demanding spatial thinking task seemed to be characterized by larger modes. Supposably, the fixation tasks demand not simply less cognitive load, but qualitatively different, which might be detected by these low-factor-IPA values. The most appropriate distribution could be found for threshold factor 0.8 with a substantial increase of the IPA values from the fixation tasks to the easiest level and stability over the following difficulty levels. All other threshold factors contain curves in their distributions. This distribution showed also the highest similarity with the distributions of the ICA values. In contrast to the ICA, the IPA showed no evidence for dependency of fatigue.

These results demonstrate that the IPA is able to differentiate between different cognitive loads. However, different threshold factors produce different distributions that are only appropriate for different regions of cognitive load. Although some results were not intuitive, this study demonstrates that the IPA (and maybe pupillary measures in general) are appropriate indicators of cognitive load but with different validity depending on the specific tasks. Especially the found dependency between the validity of the pupillary based measures and the considered cognitive task shows the importance to adapt the used algorithm with respect to the utilized stimulus materials. Hence, an appropriate pupillary based algorithm should allow adaptions and should be therefore completely fully documented, such as the IPA.

Future studies should consider different stimulus materials and should test further changes of other parameters of the IPA algorithm. The optimal parameter configuration for a specific test would depend on the criterion that the IPA is intended to indicate.

REFERENCES

John L. Andreassi. 2007. *Psychophysiology: Human behavior and physiological response* (5th ed. ed.). L. Erlbaum Publishers, Mahwah, N.J and London.

John B. Carroll. 1993. *Human cognitive abilities.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511571312

Siyuan Chen and Julien Epps. 2014. Efficient and robust pupil size and blink estimation from near-field video sequences for human-machine interaction. *IEEE transactions on cybernetics* 44, 12 (2014), 2356–2367. https://doi.org/10.1109/TCYB.2014.2306916

Benjamin Cowley, Marco Filetti, Kristian Lukander, Jari Torniainen, Andreas Henelius, Lauri Ahonen, Oswald Barral, Ilkka Kosunen, Teppo Valtonen, Minna Huotilainen, Niklas Ravaja, and Giulio Jacucci. 2015. The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction. *FNT in Human–Computer Interaction (Foundations and Trends in Human–Computer Interaction)* 9, 3-4 (2015), 151–308. https://doi.org/10.1561/1100000065

V. Demberg, K. Evangelia, and A. Sayeed. 2013. The Index of Cognitive Activity as a Measure of Linguistic Processing. In *Cooperative minds*, Markus Knauff, Michael Pauen, Natalie Sebanz, and Ipke Wachsmuth (Eds.). Cognitive Science Soc, Austin, Tex., 2148–2153.

Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The Index of Pupillary Activity. In *CHI 2018*, Regan Mandryk, Mark Hancock, Mark Perry, and Anna Cox (Eds.). The Association for Computing Machinery, New York, New York, 1–13. https://doi.org/10.1145/3173574.3173856

EyeTracking Inc. 2015. Workload RT (1.5) [Software]. http://www.eyetracking.com

Benedict C. O. F. Fehringer. 2019. *The diagnostic potential of eye tracking and pupilometry in the context of spatial thinking: Doctoral thesis, University of Mannheim.* Mannheim.

Marcel Adam Just and Patricia A. Carpenter. 1993. The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 47, 2 (1993), 310–339. https://doi.org/10.1037/h0078820

Daniel Kahneman. 1973. *Attention and effort.* Prentice Hall, Englewood Cliffs.

Sandra P. Marshall. 2000. Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity.

Sandra P. Marshall. 2009. What the Eyes Reveal: Measuring the Cognitive Workload of Teams. In *Digital human modeling*, Vincent G. Duffy (Ed.). Lecture Notes in Computer Science, Vol. 5620. Springer, Berlin and Heidelberg, 265–274. https://doi.org/10.1007/978-3-642-02809-0{_}29

Allan Paivio. 2014. Intelligence, dual coding theory, and the brain. *Intelligence* 47 (2014), 141–158. https://doi.org/10.1016/j.intell.2014.09.002

Oskar Palinko and Andrew L. Kun. 2012. Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, Carlos H. Morimoto (Ed.). ACM, New York, NY, 413. https://doi.org/10.1145/2168556.2168650

Felix D. Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological methods* 22, 2 (2017), 322–339. https://doi.org/10.1037/met0000061

Maximilian Schwalm, Andreas Keinath, and Hubert D. Zimmer. 2008. Pupillometry as a method for measuring mental workload within a simulated driving task. In *Human factors for assistance and automation*, Dick de Waard (Ed.). Shaker Publ, Maastricht, 1–13.

Stuart R. Steinhauer, Greg J. Siegle, Ruth Condray, and Misha Pless. 2004. Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 52, 1 (2004), 77–86. https://doi.org/10.1016/j.ijpsycho.2003.12.005

Adriana A. Zekveld and Sophia E. Kramer. 2014. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology* 51, 3 (2014), 277–284. https://doi.org/10.1111/psyp.12151

Bin Zheng, Xianta Jiang, and M. Stella Atkins. 2015. Detection of Changes in Surgical Difficulty: Evidence From Pupil Responses. *Surgical innovation* 22, 6 (2015), 629–635. https://doi.org/10.1177/1553350615573582