Pupillometry as a tool to study expertise in medicine

Adam Szulewski[a], Danielle Kelton[b], Daniel Howes[c]

[a]Department of Emergency Medicine, Queen's University, Canada

[b]Faculty of Medicine, Queen's University, Canada

[c]Departments of Emergency Medicine and Critical Care Medicine, Queen's University, Canada

Abstract

Background *Pupillometry has been studied as a physiological marker for quantifying cognitive load since the early 1960s. It has been established that small changes in pupillary size can provide an index of the cognitive load of an individual as he/she performs a mental task. The utility of pupillometry as a measure of expertise is less well established, although recent research in the fields of education, medicine and psychology indicates that differences in pupillary size during domain-specific tasks allows differentiation between experts and novices in appropriately designed experiments.* **Purpose** *The goal of this review is to explore the existing body of evidence for the use of pupillometry as a measure of expertise and to identify its strengths and constraints within the context of expertise research in the medical sciences.* **Results** *Pupillometry is a robust metric that allows researchers to better understand cognitive load in medical practitioners with varying levels of expertise. In medical expertise research, it has been used to study surgeons, anesthetists and emergency physicians. Its strengths include its ability to provide quantitative and objective outputs, to be measured unobtrusively with new technology and to be precisely computed as cognitive load changes over the course of completion of a task. Constraints associated with this methodology include its potential inaccuracy with changes in ambient light and pupillary accommodation as well as the need for relatively expensive equipment.* **Conclusion** *With recent technological advances, pupillometry has become a simple and robust method for quantifying physiological changes attributable to cognitive load and is increasingly being utilized in medical education. It can be used as a reliable marker of cognitive load and has been shown to differentiate levels of expertise in medical practitioners.*

Keywords: pupillometry; cognitive load; expertise; medical education

Corresponding author: ass. Prof. *Adam Szulewski BSc, MD, FRCPC, MHPE, Dep. of Emergency Medicine, Queen's University, Empire 3, Kingston General Hospital, 76 Stuart Street, Kingston, Ontario, K7L 2V7,Canada, aszulewski@qmed.ca DOI:* http://dx.doi.org/10.14786/flr.v5i3.256

1. Background

The measurement of human cognitive load has been of interest to researchers for decades. Knowing *how intensely* a person is thinking has implications beyond knowing *what* that person is thinking about. This is particularly relevant in the context of professional domains (like medicine) where critical and cognitively loading decisions often need to be made with limited time and in the context of other competing priorities.

This "intensity" of thinking, which is related to cognitive load, functions within the constraints of a limited working memory. Working memory is a key executive function. Executive functions are a group of mental processes that are required when an individual has to pay attention, and when it would be considered inappropriate, insufficient or impossible to rely on instinct or to respond automatically (Burgess & Simons, 2005). In addition to working memory, executive functioning involves two other core activities: inhibition (self-control, selective attention, cognitive inhibition) and cognitive flexibility (which is closely related to creativity). These three core activities are combined in different ways to build higher order executive functions such as reasoning, problem solving and planning (Diamond, 2013).

Working memory, which is responsible for the manipulation of stored information (or our ability to "think"), is generally thought to be limited (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). Ongoing work in this area now suggests that with experience, experts are able to expand working memory capacity by having developed methods for storage and retrieval of domain-specific information in long-term memory – so called long-term working memory (Ericsson & Kintsch, 1995). This is accomplished, in part, by pattern recognition and schema development, and a resultant relative decrease in the cognitive load that a problem or situation imposes as an individual becomes more expert-like (Szulewski, Roth, & Howes, 2015).

Functionally, cognitive load can be thought of as the mental capacity that is allocated to performing a task (Paas et al., 2003). It is thought to be comprised of three components: intrinsic cognitive load (ICL), extraneous cognitive load (ECL) and germane cognitive load (GCL) (Young, Van Merrienboer, Durning, & Ten Cate, 2014). ICL is a function of expertise and task complexity, while ECL is related to suboptimal information presentation conditions. GCL refers to the working memory resources that are dedicated to processing ICL, and thus to learning (Sweller, 2010).

In general, researchers measure cognitive load using psychometric scales, physiological variables and secondary task methodology (Paas et al., 2003). Briefly, psychometric scales gather subjective data from participant self-reports after task completion. Physiological variables use task-evoked pupillary responses (TEPRs) or pupillometry, heart rate variability, galvanic skin response (among others) as surrogate markers of cognitive load. Secondary task methodology relies on participants' performance on a secondary task (that requires sustained attention, like detecting an auditory signal) and uses this information to glean the level of cognitive load imposed by the primary task. Each of these techniques has its own strengths and limitations; but in general, each is thought to provide data about total (or measurable) cognitive load. The contribution of intrinsic, extraneous and germane cognitive load to each of these measurement techniques remains to be elucidated (Leppink, Paas, van Gog, van der Vleuten, & van Merriënboer, 2014).

This review focuses on one particular physiological method of measuring cognitive load – pupillometry, which is the study of changes in pupil size. We will first examine the technique of pupillometry as a surrogate marker for cognitive load in non-medical domains and then we will focus the discussion on pupillometry research in medicine and how this relates to the development of expertise. Finally, constraints of the technique will be discussed.
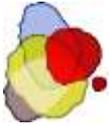
2.   Pupil physiology

Dilation and constriction of the human pupil is necessary for day-to-day visual tasks. Dilation (mydriasis) of the pupil is accomplished by the contraction of the iris dilator (radial) muscle, which is controlled by the sympathetic nervous system. Constriction (miosis) of the pupil occurs when the iris sphincter (circular) muscle contracts, which is controlled by the parasympathetic nervous system. Two commonly tested reflexes in clinical medicine are the light reflex and the accommodation reflex. During the light reflex, the pupil dilates in low luminance environments and constricts in high luminance environments. In the accommodation reflex, as an individual changes visual focus from a distant object to a closer object, the pupil constricts, and vice-versa (Lang, 2015).

In addition to these clinically measurable and commonly discussed reflexes, pupils also change in size as a result of non-visual stimuli. This was first described in detail by Hess and Polt (1960) where it was shown that pupil size varied when participants viewed particular images (for example, sexually suggestive ones). Follow-up studies by this group and others further demonstrated that pupil size could be used to measure cognitive load (or mental effort). Physiologically, it is thought that pupil size changes with cognitive loading as a result of pathways that originate in the locus coeruleus, which is a major norepinephrine source in the brain (Laeng, Sirois, & Gredebäck, 2012). In fact, locus coeruleus activity has been shown to be very closely related to sympathetic activity and changes in pupil size (Aston-Jones & Cohen, 2005). These pupillary responses are spontaneous and very difficult to control voluntarily. The voluntary dilation of a subject's pupils is only possible indirectly if the subject imagines a situation (e.g. self-induced sexual imagery) where his/her pupils would normally dilate (Whipple, Ogden, & Komisaruk, 1992). This would be particularly difficult, if not impossible, to systematically do while simultaneously performing other cognitively loading tasks. This makes the technique robust. Although other autonomic measurements like heart rate and skin resistance have also been found to provide similar information regarding sympathetic activity (and thus cognitive loading), pupillometry has been found to yield the most consistent and readily analysable results (Kahneman, Tursky, Shapiro, & Crider, 1969).

3.   Pupillometry as a measure of cognitive load

In a seminal article, Hess and Polt (1964) found there to be a strong correlation between difficulty of arithmetic problems posed to participants and the magnitude of the increase in their pupil sizes. Further, they observed that after a question was asked, participants' pupils showed a gradual increase in diameter, reached a maximum size just prior to reporting an answer, and then reverted back to their original diameter shortly thereafter. In another article, Beatty and Kahneman (1966) built upon these original experiments and were able to confirm two phases in the pupillary response to cognitive processing. First, they noted a loading phase with dilation corresponding to information gathering and an unloading phase where the pupil constricted as answers were verbalized by the participants. Based on the results from these studies as well as others, it became generally accepted that changes in pupil size reflect changes in cognitive processing load during task performance and provide information about processing resources. Specifically, more difficult cognitive tasks were found to cause both an increase in the amplitude and the latency of pupillary dilation (Beatty, 1982).

These early experiments were carried out with relatively onerous experimental processes that involved developing large quantities of photographs taken by cameras in precisely controlled environments and then manually measuring pupil size with a ruler. This made large-scale experiments impractical. Modern technology has allowed researchers to electronically collect pupil size data with stationary as well as mobile devices, obviating the need for time-consuming manual measurement and allowing for less stringent experimental environments. Some of the previously described studies that used arithmetic problems have

now been replicated with the new technology, showing similar results. Figure 1 is taken from one of these studies that used a mobile eye-tracker to capture participant pupil size at a rate of 30Hz during arithmetic problem solving. As was first shown in the original experiments, the new technology also demonstrated that pupil size increased with increasing problem difficulty and changed predictably with phases of information gathering and delivery of responses (Szulewski, Fernando, Baylis, & Howes, 2014).
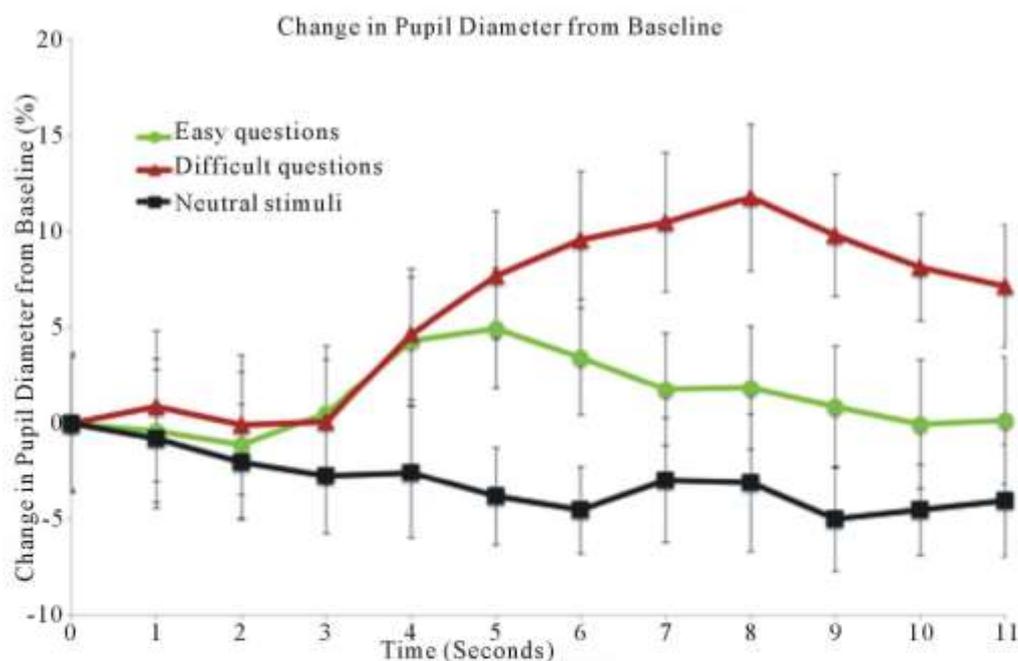

Figure 1. "Difficult questions resulted in peak dilation of 11.8% compared to baseline whereas "easy" questions resulted in peak dilation of 5.0% compared to baseline (p = 0.005). Time 0 to 3 seconds serves as baseline (3 seconds prior to question presentation); time 3 to 8 seconds corresponds to the time that the question was on the screen; time 8 to 11 seconds corresponds to the 3 seconds after the question was removed and the black dot appeared. [From Szulewski, A., Fernando, S. M., Baylis, J., & Howes, D. (2014). Increasing pupil size is associated with increasing cognitive processing demands: A pilot study using a mobile eye-tracking device. Open Journal of Emergency Medicine, 2014. Reprinted with permission].

The ease of use and precision of the newer technology has expanded the role of pupillometry to more theoretical realms. In addition to reliably demonstrating increased cognitive load with increasing question difficulty, pupillometry data have also shown that the modality of information presentation has cognitive loading effects. Using a remote eye tracker, Klingner, Tversky, and Hanrahan (2011) showed that cognitive load is higher for the same tasks when they are presented orally as opposed to visually. These experiments underscore the precision and expanded applications of this technique.

Other groups of researchers have also investigated the ability to measure cognitive load in novel environments using pupillometry. One such experiment by Palinko, Kun, Shyrokov, and Heeman (2010) investigated measuring mean pupil diameter change in drivers as they operated a simulated vehicle while they were involved in simultaneous spoken dialogues. Pupil diameter changed as expected and the authors concluded that pupillometry was better in quantifying small changes in cognitive load in the simulator compared with other measures like lane position and steering wheel angle. Results from studies like this one suggest that pupillometry can be reliably used in more true-to-life situations in addition to well controlled laboratory settings. Importantly, during the driving simulator experiment, luminance varied only ± 5% in the simulated experimental environment which likely minimized the contribution of the light reflex to pupillary

changes and allowed for a relatively clean signal. Changes in luminance become more of an experimental issue in real-world environments where background luminance varies to greater degrees.

On the whole, these studies seem to suggest that the construct being measured with pupillometry is cognitive load, although there is research to suggest that other factors (e.g. emotion, fatigue, age, pain and certain drugs) also contribute to change in pupil size (Holmqvist et al., 2011). Validity evidence for the use of pupillometry to measure cognitive load specifically was recently described by Szulewski, Gegenfurtner, Howes, Sivilotti, and van Merriënboer (2016). In this study, pupillometric measurements of cognitive load were compared to psychometric measurements of cognitive load across different question types, question difficulty and experience levels in a testing environment. Based on the predictability of the results and the strong correlation of the measurement instruments, the authors concluded that there is validity evidence to use either psychometric or pupillometric measurements to measure cognitive load in traditional testing environments.

4. Pupillometry research in medicine

Given the promising results of pupillary analysis in experimental settings and the increasing availability of the new technology, researchers have started to expand its use into other domains, including medicine. Medicine is a particularly interesting field in which to study cognitive load given its inherent characteristics where physicians regularly make high stakes decisions, often under considerable external pressures (including time and stress).

These characteristics are emphasized during non-routine emergency situations. One study that investigated critical incidents in the operating room examined anaesthesiology trainees' pupil sizes, among other physiological responses as surrogate markers of cognitive workload (Schulz et al., 2011). Participants' pupil sizes were found to increase as the severity of a critical incident increased. Although this pattern held true within scenarios, the authors found that there was no difference between sessions or individuals. This was thought to be due to individual pupil variations as well as external factors like lighting.

These issues raise the concern that external factors can skew pupillometric results and make it difficult to interpret the data reliably in real-world environments where luminance is not adequately controlled. All real-life physician-patient clinical encounters would, as a result, be affected. The main issue in these scenarios involves the light reflex which is capable of causing pupil diameter changes of up to 120% from baseline, which is far greater than the changes of up to 20% that can be attributed to cognitive processing demands. (Holmqvist et al., 2011; Laeng et al., 2012).

In an effort to mitigate the confounding effects of the light reflex, investigators often try to control for luminance during their experiments. Zheng, Jiang, and Atkins (2015) did just this and were able to confirm that pupil responses behaved as expected with changing sub-task difficulty in a simulated laparascopic surgical experiment. In a related study, the same group noted that the rate of change of pupil size was better than pupil diameter in assessing mental workload of the simulated laparscopic task (Jiang, Zheng, Tien, & Atkins, 2013).

To address some of these issues, new techniques (like the index of cognitive activity) have been designed to separate out the light reflex from pupil changes secondary to cognitive workload by measuring abrupt discontinuities in the pupil size signal (Marshall, 2002). This index of cognitive activity has been utilized in the objective assessment of surgical skill where pupil size (along with other eye and pupillary metrics) was used to objectively classify non-expert from expert surgeons in environments that were uncontrolled for luminance including a simulator as well as a live operating room (Richstone et al., 2010).
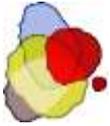
5.   Pupillometry and expertise

Performance on tests has universally been utilized to measure the construct of ability, intelligence, competence or expertise in a domain. Despite its wide use, test-taking is known to have many limitations as a surrogate marker to measure these constructs. Applications of pupillometry have allowed researchers to delve deeper into this area than simply examining performance. This is particularly interesting when considering cognitive processing as subjects answer questions correctly. Based on the traditional view of assessment in test-taking, two individuals who get the same score on a test are thought to have equal domain-specific skill (or ability or expertise). The reality is more nuanced. Figure 2 is taken from a study by Ahern and Beatty (1979) which shows the cognitive processing demands (as measured by pupillometry) of participants with both high and low intelligence as defined by Scholastic Aptitude Test scores as they were faced with arithmetic problems (and answered these problems correctly). Based on traditional assessment modalities, both groups of individuals would be assessed equally for having answered correctly. A closer analysis, however, revealed that the group with "lower intelligence" had greater increases in pupillary dilation than the "higher intelligence" group at all question difficulty levels. Essentially, the group with the lower intelligence had to "think harder" to achieve the same correct response as the group with higher intelligence.
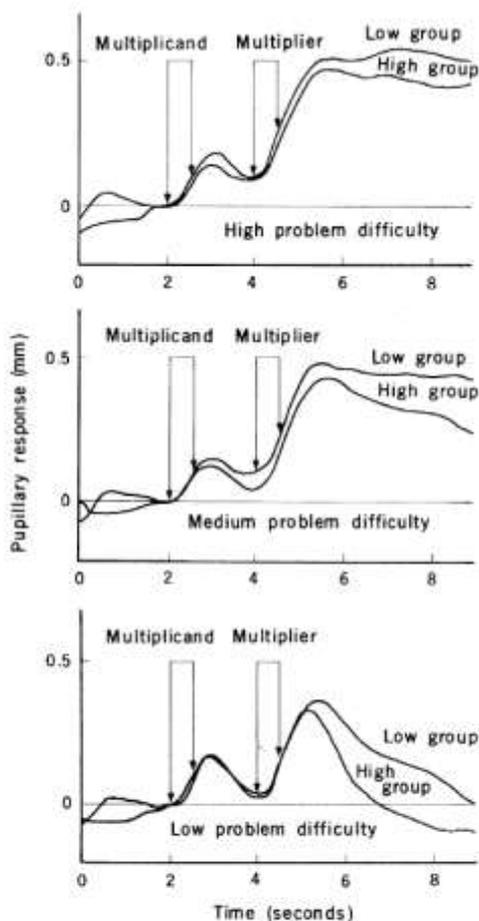

Figure 2. Averaged task-evoked pupillary responses for correctly solved problems at three levels of difficulty for subjects in the high and low groups of psychometrically measured intelligence. At all difficulty levels, larger pupillary responses are observed for subjects in the low group. [From Ahern, S., & Beatty, J.

(1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205*(4412), 1289-1292. Reprinted with permission from AAAS.]

Moving away from intelligence defined by standardized testing, a study by Szulewski et al. (2015) found similar results in novices and trained physicians as they answered clinically-based multiple choice questions. The participant groups in this study were divided not by intelligence, but by clinical experience. Those with more clinical experience (the trained physician group) had smaller changes in pupil diameter as they answered the questions compared to the more novice group when both groups answered correctly (See Figure 3). In another study, Tien et al. (2015) found that junior surgeons had greater pupil sizes than expert surgeons during open inguinal hernia repair. Both of these studies (which divided physician participant groups based on experience level) emphasize that those with less experience expend more cognitive load than those with more experience when they perform domain-specific tasks, even when the measured outcome is the same.

Although it is reasonable to assume that these observed differences are due to different experience levels, one might argue that there may be other confounding factors between the groups that could skew the results. This is a potential issue for any cross-sectional study. A study by Richstone et al. (2010) suggests that it is in fact experience/expertise that is responsible for the pupillometric changes between groups, as opposed to another confounder. As part of their study, they examined one non-expert surgeon three times over the course of 18 months both in simulated and live surgical environments. During this longitudinal analysis, they found that it became increasingly difficult to differentiate this non-expert from the expert surgeon group as his pupil metrics became more expert-like over time with increased training and experience. This finding suggests that the differences between groups of participants are in fact due to skill or expertise, as opposed to another confounding factor. Overall, these studies suggest that there is empirical evidence that those with more domain-specific experience exhibit a certain cognitive efficiency as they perform tasks associated with their training and experience that their novice counterparts have not yet developed.
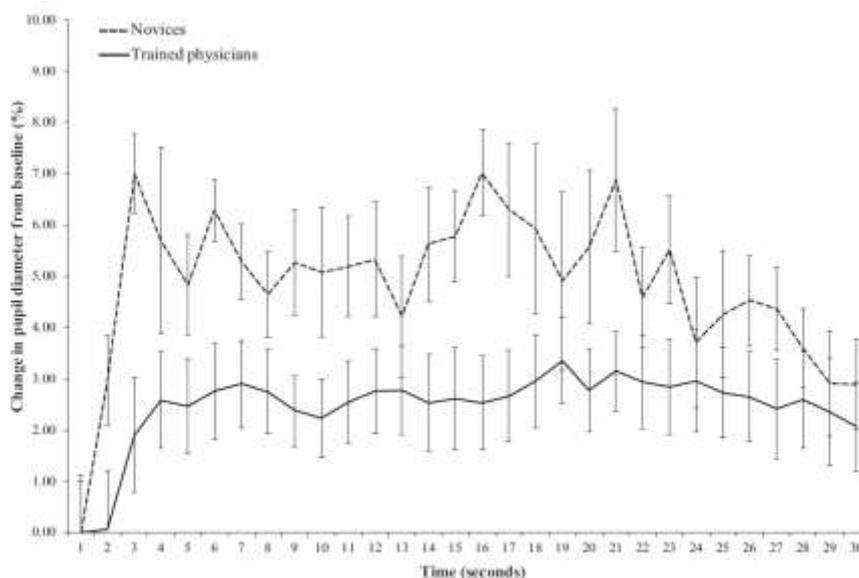

Figure 3. Results of an analysis of correctly answered clinical multiple-choice questions. The increase in the pupil diameter of novices was significantly greater than that of trained physicians ($P < 0.001$). [From Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987. Reprinted with permission from AAMC.]
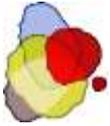
It is debatable whether differences in cognitive efficiency are relevant during a test where a student is asked a sequence of questions and he/she generally focuses all of his/her working memory onto the question at hand before moving onto the next one. Moreover, it is debatable whether an assessor would even want to know this information. Arguably, however, this cognitive efficiency in the "more intelligent" or "more skilled" or "more experienced" or "more expert-like" group becomes relevant in complex situations with competing priorities, as the less cognitively strained individual will have a greater proportion of his/her working memory available for other cognitively demanding executive functions.

One such area where competing priorities often coexist and where cognitive efficiency might be beneficial is clinical medicine. During medical emergencies in particular, a physician team leader is cognitively tasked not only with making appropriate medical decisions but also employing crisis resource management techniques (like leadership skills, situational awareness, communication skills and resource utilization) to optimize patient care (Hicks, Bandiera, & Denny, 2008). It logically follows that cognitive efficiency in medical decision-making will more readily allow the physician leader to perform these simultaneous crisis resource management tasks to a higher level given the real constraints of human working memory. Anecdotally, cognitive efficiency seems to evolve with experience. The "anatomy" of working memory is thought to change with the development of expertise and it is likely that certain clinical tasks cognitively load experts and novices in different ways (Szulewski et al., 2015). This evolution of the thinking process is tied to expertise development.

6.   Typical experimental conditions for pupillometry studies

As outlined in this review, researchers have successfully used pupillometry as a cognitive load measurement tool in numerous experimental conditions. These range from relatively simple experiments where pupillometric data are gathered as participants are presented with written or verbal questions and are tasked to solve problems in fields including arithmetic and language, among others. Other experiments involve the use of different stimuli including photographs or even simulated driving environments. In medical applications, pupillometry has similarly been used in various settings including test-taking as well as more high fidelity environments like simulation and actual physician-patient clinical encounters. The task instructions provided to participants are equally variable and range from solving provided problems to performing operations in live surgical environments.

7.   Constraints of pupillometry

Though it is clear that pupillometry provides useful information about both visual as well as non-visual stimuli, the technology has a number of constraints. Until relatively recently, accurate pupillometry studies required cumbersome experimental environments and tedious data collection and analysis. Although some of these issues have been addressed with new technology, the cost of this technology poses new financial barriers for certain researchers on smaller budgets. This is especially relevant for those part-time researchers who may want to incorporate pupillometry into their professional and teaching duties, like academic physicians. This reality suggests that, for the time being, given the costs, pupillometry research is more likely to occur at a theory-building level. As a result, some of its potential benefits in adjusting task difficulty for an individual learner and individualizing and optimizing education will remain elusive until the technology becomes cheaper and more readily available for teachers.

Another significant constraint of the technology relates to the accommodation and light reflexes. Although pupillometry provides consistent and fairly easy-to-interpret data in experimental conditions of constant ambient light and focus distance, data output in real-world conditions is suboptimal. As previously discussed in this review, the index of cognitive activity has been designed in an effort to overcome some of these obstacles. Although this technique allows for extraction of valuable information from large data sets with changing ambient light, the results are more coarse and provide less precise and detailed information about shorter-term cognitive changes that might be relevant in studying precise comparisons between groups performing shorter tasks (Klingner, Kumar, & Hanrahan, 2008). In addition, because this metric is a commercial product and its algorithm is not made publicly available, it cannot be replicated nor adequately studied. As a result, it is of limited benefit to researchers.

Another consideration in pupillometry research is participant age. Older individuals generally have pupils that are smaller and are more restricted in their ability to dilate compared to younger people (Holmqvist et al., 2011; Piquado, Isaacowitz, & Wingfield, 2010). Since many studies compare cognitive load between novices (who are usually younger) to experts (who are generally older), this might lead to confounding, as a smaller pupil diameter change may be due to a combination of increased age as well as decreased cognitive load. Cognitive load researchers should be aware of this issue and either control for participant age (where possible) or correct for it. Correction measures include expressing pupil size changes relative to a baseline measurement and/or age-adjusting for pupil size and reactivity based on participants' pupil responsiveness to a range of experimental light stimuli (Piquado et al., 2010).

Finally, the accuracy of pupillometric measurement is dependent to some degree on gaze position, with greater systematic error occurring when the eye is looking away from the eye-tracker's camera (Brisson et al., 2013). Different eye-tracking devices attempt to correct for this error, but the accuracy of pupillometric measures suffers from variable quality under these conditions. This is especially relevant for researchers studying cognitive load where the participant's gaze may move away from centre.

8.   Conclusion

Pupillometry is a robust and reliable method for studying cognitive load. Since its inception as a scientific field in the 1960's, it has evolved greatly. The development of new technology to measure pupil size that can electronically gather pupil data at high rates has led to the increased use of pupillometry in diverse fields. Despite the inherent constraints of the technique including interference by luminance and its cost, pupillometry remains a promising metric for researchers to utilize in the study of cognitive load. It can provide insights into the human thinking process that would otherwise be unobservable. It has a particularly promising role in the field of medicine and in the study of physician expertise development. Utilizing pupillometry to better understand and optimize physician cognitive load (and overload) is clinically relevant and has the potential to directly impact medical education and ultimately patient care.

Keypoints

- Pupillometry is a robust method of quantifying cognitive load.

- Otherwise unobservable insights into cognitive processes can be gleaned with the use of pupillometry.

- Pupillometry research in medicine is contributing to a better understanding of expertise development across medical domains.

Despite its benefits, pupillometry data in real-world applications suffers in quality as a result of the light and accommodation reflexes.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205*(4412), 1289-1292.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci., 28*, 403-450. doi: 10.1146/annurev.neuro.28.061604.135709

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin, 91*(2), 276.

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science, 5*(10), 371-372.

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods, 45*(4), 1322-1331. doi:10.3758/s13428-013-0327-0

Burgess, P. W., & Simons, J. S. (2005). Theories of frontal lobe executive function: clinical application. In P. W. Halligan & D. T. Wade (Eds.), *Effecitveness of Rehabiliation for Cognitive Deficits* (pp. 211-231). New York: Oxford Univeristy Press.

Diamond, A. (2013). Executive Functions. *Annual review of psychology, 64*, 135-168. doi:10.1146/annurev-psych-113011-143750

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological review, 102*(2), 211.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132*(3423), 349-350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Hicks, C. M., Bandiera, G. W., & Denny, C. J. (2008). Building a Simulation- based Crisis Resource Management Course for Emergency Medicine, Phase 1: Results from an Interdisciplinary Needs Assessment Survey. *Academic Emergency Medicine, 15*(11), 1136-1143. doi: 10.1111/j.1553-2712.2008.00185.x

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*: OUP Oxford.

Jiang, X., Zheng, B., Tien, G., & Atkins, M. (2013). Pupil response to precision in surgical task execution. *Studies in health technology and informatics, 184*, 210.

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology, 79*(1p1), 164.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). *Measuring the task-evoked pupillary response with a remote eye tracker*. Paper presented at the Proceedings of the 2008 symposium on Eye tracking research & applications.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48*(3), 323-332. doi: 10.1111/j.1469-8986.2010.01069.x

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on psychological science, 7*(1), 18-27. doi: 10.1177/1745691611427305

Lang, G. K. (2015). *Ophthalmology*: Thieme.

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P., & van Merriënboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32-42.

Marshall, S. P. (2002). *The index of cognitive activity: Measuring cognitive workload.* Paper presented at the Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist, 38*(1), 63-71. doi: 10.1207/S15326985EP3801_8

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). *Estimating cognitive load using remote eye tracking in a driving simulator.* Paper presented at the Proceedings of the 2010 symposium on eye-tracking research & applications.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560-569. doi:10.1111/j.1469-8986.2009.00947.x

Richstone, L., Schwartz, M. J., Seideman, C., Cadeddu, J., Marshall, S., & Kavoussi, L. R. (2010). Eye metrics as an objective assessment of surgical skill. *Annals of surgery, 252*(1), 177-182. doi: 10.1097/SLA.0b013e3181e464fb

Schulz, C., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Wasmaier, M., . . . Schneider, G. (2011). Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment. *British journal of anaesthesia, 106*(1), 44-50. doi: 10.1093/bja/aeq307

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123-138. doi: 10.1007/s10648-010-9128-5

Szulewski, A., Fernando, S. M., Baylis, J., & Howes, D. (2014). Increasing pupil size is associated with increasing cognitive processing demands: A pilot study using a mobile eye-tracking device. *Open Journal of Emergency Medicine, 2014*. doi: 10.4236/ojem.2014.21002

Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L. A., & van Merriënboer, J. J. G. (2016). Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, 1-18. doi: 10.1007/s10459-016-9725-2

Szulewski, A., Roth, N., & Howes, D. (2015). The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *Academic Medicine, 90*(7), 981-987. doi: 10.1097/ACM.0000000000000677

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G.-Z., & Darzi, A. (2015). Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. *Surgical endoscopy, 29*(2), 405-413. doi: 10.1007/s00464-014-3683-7

Whipple, B., Ogden, G., & Komisaruk, B. R. (1992). Physiological correlates of imagery-induced orgasm in women. *Archives of sexual behavior, 21*(2), 121-133.

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: Implications for medical education: AMEE guide no. 86. *Medical teacher, 36*(5), 371-384. doi: 10.3109/0142159X.2014.889290

Zheng, B., Jiang, X., & Atkins, M. S. (2015). Detection of Changes in Surgical Difficulty: Evidence From Pupil Responses. *Surgical Innovation, 22*(6), 629-635. doi: 10.1177/1553350615573582