Measuring physician cognitive load

Adv in Health Sci Educ (2017) 22:951–968
DOI 10.1007/s10459-016-9725-2

Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool

Adam Szulewski[1] · Andreas Gegenfurtner[2] · Daniel W. Howes[3] · Marco L. A. Sivilotti[4] · Jeroen J. G. van Merriënboer[5]

Received: 25 February 2016 / Accepted: 19 October 2016 / Published online: 27 October 2016

Abstract In general, researchers attempt to quantify cognitive load using physiologic and psychometric measures. Although the construct measured by both of these metrics is thought to represent overall cognitive load, there is a paucity of studies that compares these techniques to one another. The authors compared data obtained from one physiologic tool (pupillometry) to one psychometric tool (Paas scale) to explore whether they actually measured the construct of cognitive load as purported. Thirty-two participants with a range of resuscitation medicine experience and expertise completed resuscitation-medicine based multiple-choice-questions as well as arithmetic questions. Cognitive load, as measured by

✉ Daniel W. Howes
howesd@kgh.kari.net

Adam Szulewski
aszulewski@qmed.ca

Andreas Gegenfurtner
andreas.gegenfurtner@th-deg.de

Marco L. A. Sivilotti
marco.sivilotti@queensu.ca

Jeroen J. G. van Merriënboer
j.vanmerrienboer@educ.unimaas.nl

[1] Department of Emergency Medicine, Queen's University, Kingston General Hospital, 76 Stuart Street, Kingston, ON K7L 2V7, Canada

[2] Institut für Qualität und Weiterbildung, Technische Hochschule Deggendorf, Edlmairstraße 9, 94469 Deggendorf, Germany

[3] Departments of Emergency Medicine and Critical Care, Queen's University, Kingston General Hospital, 76 Stuart Street, Kingston, ON K7L 2V7, Canada

[4] Departments of Emergency Medicine and Biomedical and Molecular Sciences, Kingston General Hospital, 76 Stuart Street, Kingston, ON K7L 2V7, Canada

[5] Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Universiteitssingel 60, Room N.5.06, 6200 MD Maastricht, The Netherlands

both tools, was found to be higher for the more difficult questions as well as for questions that were answered incorrectly ($p < 0.001$). The group with the least medical experience had higher cognitive load than both the intermediate and experienced groups when answering domain-specific questions ($p = 0.023$ and $p = 0.003$ respectively for the physiologic tool; $p = 0.006$ and $p < 0.001$ respectively for the psychometric tool). There was a strong positive correlation (Spearman's $\rho = 0.827$, $p < 0.001$ for arithmetic questions; Spearman's $\rho = 0.606$, $p < 0.001$ for medical questions) between the two cognitive load measurement tools. These findings support the validity argument that both physiologic and psychometric metrics measure the construct of cognitive load.

Keywords Cognitive load · Expertise · Eye-tracking · Psychometrics · Pupillometry · Resuscitation · Validity

Introduction

Physician cognitive load is an intrinsic characteristic of work in acute-care medical settings and is known to affect performance (Perry et al. 2013). The nature of work in the emergency department, where physicians are frequently interrupted, treat multiple patients simultaneously and must regularly prioritize decision-making, places considerable demands on their cognitive resources and thus increases the likelihood of making errors (Laxmisan et al. 2007). To greater or lesser degrees, these observations hold true in many domains of medicine where physicians must balance competing priorities while caring for their patients.

From a theoretical perspective, cognitive load is thought to be comprised of three basic elements: intrinsic cognitive load, extraneous cognitive load and germane cognitive load (Young et al. 2014). Intrinsic cognitive load is a function of the complexity of the information to be processed and the expertise of the task performer; while extraneous cognitive load is due to suboptimal information presentation conditions. The sum of intrinsic and extraneous cognitive load is thought to represent the overall cognitive load that can be measured experimentally. Germane cognitive load is thought to refer to the working memory resources dedicated to actively processing intrinsic cognitive load, and thus to learning (Sweller 2010).

Cognitive load theory (CLT) is a theory of learning based on the optimal design of instructional methods that considers a learner's finite cognitive capacities to apply knowledge and transfer it to new situations (Paas et al. 2003). According to CLT, mental processing is limited by the capacity of working memory (De Jong 2010; Sweller et al. 1998). With the development of domain-specific expertise, those with more experience are thought to be able to chunk related concepts together in elaborated schemas, thus maximizing the efficiency of their working memory (Gegenfurtner et al. 2011; Sweller et al. 1998). In addition to the efficiency afforded by schema creation, individuals are thought to be able to extend domain-specific long-term working memory with experience in a given field despite the traditional supposition that working memory itself is static (Ericsson and Kintsch 1995). In medicine, this is accomplished through the development of retrieval cues between working memory and long-term memory that accelerate memory encoding and decoding. Richer mental models are created which allow experienced clinicians to more readily recognize when a new clinical scenario may fit with a previously identified pattern

(Gegenfurtner and Seppänen 2013). These same skills allow clinicians to efficiently recognize when a new clinical scenario might not fit a previously identified mental model, thus altering subsequent management decisions (Schubert et al. 2013). Deliberately practicing these cognitive strategies (as well as others) in the context of years of experience allows certain individuals to develop expertise in a domain (Ericsson et al. 2007; Norman 2005). As a result, a particular task may yield high intrinsic cognitive load for a novice task performer but a much lower intrinsic load for an expert task performer.

For decades, researchers have been interested in measuring cognitive load because it impacts the understanding of expertise development as well as education. It has been shown that measures of cognitive load can reveal important information about CLT beyond traditional performance metrics (Paas et al. 2003). The science of cognitive load quantification has been traditionally separated into physiologic measurements and psychometric measurements of this construct (Paas et al. 2003). Dual-task performance techniques (which are based on the premise that limited cognitive resources exist that must be distributed between two competing tasks) have gained some popularity in the literature as a means to quantify cognitive load as well (Brunken et al. 2003).

A well-studied method for physiologic measurement of cognitive load is pupillometry. Pupillometry consists of recording a participant's changes in pupil diameter as he/she utilizes cognitive resources for working memory processes. Pupil diameter increases as cognitive load increases as a result of central autonomic nervous system activity. As such, pupillometry is thought to provide an estimate of the intensity of a participant's cognitive load at a given instant in time (Laeng et al. 2012). Numerous studies in various fields have also found pupillometry to be useful to measure cognitive load (Beatty 1982; Hess 1965; Hess and Polt 1964; Kahneman and Beatty 1966; Klingner et al. 2008, 2011; Paas et al. 2003; Szulewski et al. 2014).

With respect to resuscitation medicine content and resuscitation medicine expertise, measuring changes in pupil size as a surrogate marker for cognitive load has shown that experienced physicians expend less cognitive load when answering domain-specific multiple choice questions than novices (Szulewski et al. 2015). It is postulated that experts' lower level of cognitive load in a testing environment is related, in part, to their expertise in authentic clinical situations (like work in an emergency department) and their expanded long-term working memory.

In addition to physiologic measures of cognitive load, psychometric scales that measure subjective cognitive load are widely used in the literature. One such example is the nine-point mental effort scale developed by Paas (1992). This scale has been widely used in the literature and has been shown to be a reliable and valid measure of overall cognitive load (Ayres 2006; Paas et al. 2003). A copy of this scale is included in "Appendix 1".

Though both physiologic and psychometric measures attempt to quantify cognitive load, there is no accepted gold-standard for cognitive load measurement. Some authors have questioned whether data derived from psychometric surveys might actually give information about intrinsic cognitive load, as opposed to overall cognitive load, as has been traditionally assumed (Naismith et al. 2015). Others have brought into question whether construct validity truly exists and if another variable, like stress, may actually be the one being measured using these techniques. A recent systematic review on cognitive load measures concluded that the quality of evidence for cognitive load measurement is low and that multiple quantification techniques should be used together in future studies to address this issue (Naismith and Cavalcanti 2015). In short, consensus about the validity of these measures does not fully exist. Moreover, there is a paucity of studies that compares physiologic and psychometric tools head-to-head. Without more evidence, it would be

premature to conclude that they are reliably measuring the same construct and that this construct is, in fact, cognitive load. This study attempts to bridge this gap, by providing evidence of validity using aspects of Cook's review of the Messick validity framework as a guide (Cook and Beckman 2006). This framework suggests that evidence to support validity of an instrument should be based on information from five sources (content, response process, internal structure, relations to other variables and consequences).

The objective of this experiment was to investigate the relationship of measured cognitive load as determined by (1) an analysis of changing pupil size and (2) responses to a subjective psychometric mental effort questionnaire. These experiments were carried out in participants with varying levels of resuscitation medicine experience as they performed a resuscitation medicine test.

Methods

Experimental setting

Participant cognitive load was measured by both physiologic and psychometric measures as participants with varying levels of resuscitation medicine experience answered a multiple-choice question (MCQ) test presented to them on a computer monitor. A research assistant who was not involved in data analysis or experimental design conducted the experiment with each participant.

Participants

A convenience sample of 32 participants was recruited between September and November of 2014. Participants were grouped according to their experience in resuscitation medicine. The novice group comprised 13 undergraduate medical students in their first two years of medical school. The intermediate group consisted of 9 senior residents (fourth or fifth year residents enrolled in emergency medicine and other resuscitation-based fields) as well as emergency medicine attending physicians in their first years of practice. The experienced group of participants included 10 attending physicians with more than ten years of clinical experience in fields related to emergency and resuscitation medicine.

The mean age of the 32 participants was 34.1 (SD = 10.8) years. The mean experience level, defined as number of years since starting medical school for all participants, was 10.1 (SD = 10.2). Participant mean age was 24.3 (SD = 1.8), 33.1 (SD = 2.6), and 47.7 (SD = 7.2) years for the novice, intermediate and experienced groups, respectively. Female participants made up 6 of the 13 novice subjects, 1 of 9 intermediate participants and 2 of 10 experienced participants. It had been a mean of 0.8 (SD = 0.4), 9.3 (SD = 2.1), and 22.9 (SD = 7.2) years since the start of medical school for the novice, intermediate and experienced groups respectively.

The rationale for this division of participants was to provide evidence of "relations with other variables" for the validity argument of the testing instruments used. *Relations with other variables evidence is* thought to bolster the validity argument when the results from subgroups based on training status vary as expected. The authors hypothesized that the novice group would have a relatively low content-knowledge, but a high test-taking ability as a result of their temporal proximity to similar MCQ testing. The intermediate group was expected to have both a high content-knowledge as well as test-taking ability. In contrast,

the experienced group was hypothesized to have a high content-knowledge but a relatively lower test-taking ability because of the increased time elapsed from their own written MCQ examinations.

Thirty-five participants were contacted by email by one of the authors to take part in the study; they did not receive an incentive to participate. Eligibility was determined based upon known training/experience level. One potential participant from the intermediate group and two potential participants from the experienced group declined to participate. All novices approached agreed to participate. The Research Ethics Board at Queen's University provided approval for this study (SMED–115-13; extension of file #6010680).

Tools used

Prior to each individual session, participants were fitted with the Tobii© Glasses Eye Tracker (Tobii Technologies, Danderyd, Sweden) and the device was calibrated as per manufacturer recommendations. During this calibration, each participant's pupil size at baseline was determined. The equipment subsequently calculated the dynamic monocular pupil size as a percentage of baseline at a rate of 30 Hz throughout the experimental session.

After answering each question, participants were prompted to rate their mental effort using the psychometric mental effort questionnaire developed by Paas (1992). See "Appendix 1" for details. Participants verbalized their responses to the questions and surveys; these audio data were recorded synchronously and analyzed later by a research assistant blinded to the pupillometry data.

These two tools were utilized in the current experiment in an effort to provide validity evidence based on "relations to other variables". This source of evidence for validity is based on the idea that if the tools are measuring the same construct, then there should be a correlation between their scores (Cook and Beckman 2006).

Furthermore, the Paas scale was used in an effort to see whether the pupillometry data was measuring what it was purported to measure based on the thought process of the participants. If the actions (pupillometry data in this case) fit with the thought process of participants, this would provide evidence validity based on "response process" (Cook and Beckman 2006). Because the Paas scale asks participants to rate their level of investment of cognitive resources and pupillometry is supposed to quantify the investment of cognitive resources, it was theorized that a correlation between the two tools would provide some of this "response process" evidence.

Instrument validity and reliability

Using various measuring devices, researchers have been using pupillary measurements as a surrogate marker for cognitive load and have validated its use in numerous experimental realms. Beatty (1982) found that digit and linguistic tasks of increasing complexity caused pupillary size to increase to greater degrees. In an early experiment, Hess (1965) showed that pupil size increased when arithmetic problems were presented to participants, peaked when the answer was given and then dropped off again. Further, higher peak pupil dilation has been shown to be associated with increasing task difficulty (Klingner et al. 2011). Szulewski et al. (2014) were able to replicate these findings using arithmetic questions with newer mobile eye-tracking technology. This new technology has also been successfully

utilized in cognitive load measurement experiments in medical testing where questions posed to participants were shown to affect pupil size in predictable ways based on question and participant characteristics (Szulewski et al. 2015). Other groups of researchers have also found consistent results using the mental effort scale developed by Paas found in "Appendix 1" (Ayres 2006; Tuovinen and Paas 2004). For example, in a group of high school students solving algebra problems, Ayres (2006) showed that the Paas scale provided a cognitive load rating that was reliable and correlated highly with errors, as expected.

Experimental design

After calibration of the eye-tracking device, each participant sat at a distance of 1 m from a computer monitor on which questions were displayed. Ambient light and screen brightness were standardized and participants were asked to continue looking at the screen throughout the experiment and to verbalize their responses to the questions.

Each participant encountered the same questions in the same order. Four arithmetic questions were interspersed among twelve resuscitation-based medical MCQ's. Questions were classified a priori into "difficult" and "easy" questions based on their origin (medical student handbook vs. specialty board examination preparation material) as well as the authors' judgement in an effort to provide evidence from a "relations to other variables" source. A black circle was presented on the monitor between questions to re-establish a pre-question pupil diameter baseline for each question and each participant.

After each question, participants were prompted to rate their mental effort using the mental effort scale in "Appendix 1" (Paas scale).

Data analysis

In general, when a participant reads a question or is presented with a problem to solve, his/her pupil diameter increases steadily until he/she provides an answer, at which point the pupil diameter decreases again (Kahneman and Beatty 1966). Previous studies utilizing pupillometry as a physiologic measure of cognitive load have concluded that measuring cognitive load accurately requires an analysis of both the magnitude of the change in pupil size as well as the duration of time between question presentation and verbalization of a response (i.e. the time that a participant is thinking about an answer) (Szulewski et al. 2015).

To combine these two parameters into one measure we calculated the area under the curve (above baseline) of the change in pupil size (expressed as a percentage of baseline pupil diameter at time of calibration) versus time from question start to verbalization of an answer [this is referred to as *pupillary change index* (PCI) throughout this manuscript and expressed in units of % seconds]. The size of this value was hypothesized to represent the participant's overall cognitive load for a given question. The determination of this value was accomplished by manual graphical analysis of the raw data to obtain a quantitative measure for each participant and each question. In order to account for possible baseline drift or residual cognitive load from a previous question, the baseline value for pupil size was recalibrated for each question by averaging the raw data just before each question was presented (corresponding to the time that each participant was focusing on a black circle presented between each question). See "Appendix 2" for a visual representation of one example of the raw pupillometric data. The manual determination of the PCI was labour-intensive (about 15 min for each PCI); accordingly, we

decided to focus this analysis on the first half of both the arithmetic and medical questions for all participants.

Peak pupil size was determined for all questions and all participants. To aid in this analysis, a procedure in Visual Basic was implemented to clean and analyze the pupillometry data (available from the authors on request). To reduce artefactual (e.g. blinking) and missing data, the 30 Hz raw data was smoothed by replacing values that were blank or deviated by more than 10 % absolute from the previous value with the rolling average of the previous 1/6 of a second.

The psychometric survey responses and the physiologic pupil data were then compared by question type, level of participant experience, correctness and level of question difficulty.

Correlational analyses were performed using parametric statistics (Pearson correlation) as well as non-parametric statistics (Spearman's ρ) as not all data were normally distributed. Other analyses were made by Pearson Chi square, student's *t* test (non-parametric Mann–Whitney U), ANOVA (non-parametric Kruskal–Wallis), and post hoc Tukey analysis. IBM SPSS Statistics 21 was used for all analyses. Correlation effect sizes were designated a priori as weak (0.10–0.29), moderate (0.30–0.49) and strong (≥0.50) (Cohen 1988). Differences were considered to be significant at a level of $p < 0.05$.

Results

Correlation between PCI and Paas scale

Overall, the PCI (a physiologic measure of cognitive load) correlated well with the Paas scale (a psychometric measure of cognitive load).

For the arithmetic questions, parametric analysis revealed a Pearson's correlation coefficient of 0.675 ($p < 0.001$), which indicates a strong positive relationship. For the medical questions, parametric analysis revealed a Pearson's correlation coefficient of 0.542 ($p < 0.001$), which also indicates a strong positive relationship. Non-parametric correlational analyses were also performed in order to confirm these values [Spearman's ρ for arithmetic questions was 0.827 ($p < 0.001$); Spearman's ρ for medical questions was 0.606 ($p < 0.001$)]. See Fig. 1 for a scatterplot of PCI values plotted against Paas scale result values. These strong correlations persisted when analyzed by participant level of experience (Table 1).

Performance based on training subgroup

Table 2 provides a summary of performance by experimental group as well as question type. There was no significant difference in performance on the arithmetic questions between the novice group and the intermediate group, the novice group and the expert group and the intermediate group and the expert group ($p = 0.858$, $p = 0.410$, and $p = 0.555$ respectively). For the medical questions, the novice group performed significantly worse than both the experienced and intermediate groups ($p < 0.001$). Though both the intermediate and experienced groups were fairly accurate, the intermediate group significantly outperformed the experienced group ($p = 0.044$).

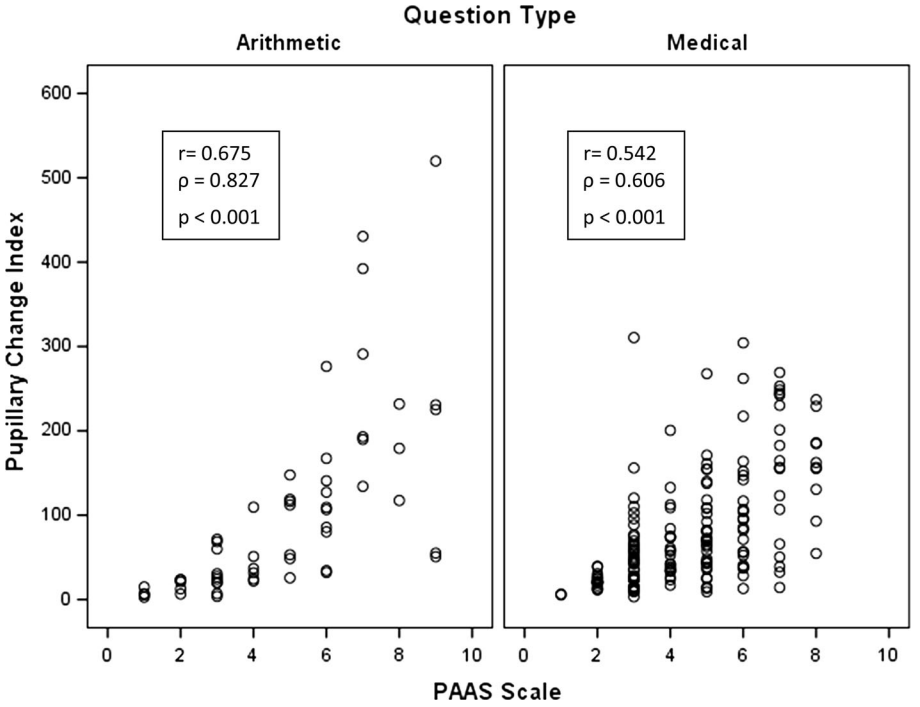


Fig. 1 Graphical representation of the correlation between pupillary change index (in % seconds) plotted against Paas scale response for all participants (novice, intermediate and experienced) (r = Pearson's r; ρ = Spearman's ρ)

Table 1 Parametric (Pearson's r) and non-parametric (Spearman's ρ) correlation coefficients of pupillary change index versus Paas scale response for the arithmetic and medical questions, broken down by subgroup

	Arithmetic questions		Medical questions	
	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
Novice	0.644 ($p = 0.003$)	0.716 ($p = 0.001$)	0.422 ($p = 0.001$)	0.501 ($p < 0.001$)
Intermediate	0.807 ($p < 0.001$)	0.834 ($p < 0.001$)	0.561 ($p < 0.001$)	0.607 ($p < 0.001$)
Experienced	0.636 ($p = 0.003$)	0.851 ($p < 0.001$)	0.639 ($p < 0.001$)	0.611 ($p < 0.001$)
Overall	0.675 ($p < 0.001$)	0.827 ($p < 0.001$)	0.542 ($p < 0.001$)	0.606 ($p < 0.001$)

Table 2 Mean (95 % confidence interval) proportion of questions answered correctly by subgroup

	Arithmetic questions (%)	Medical questions (%)
Novice	73 (59–93)	32 (25–40
Intermediate	74 (58–86	89 (82–94
Experienced	80 (65–90	79 (71–85

Additional evidence for PCI

Variation across training subgroup and question type

Arithmetic questions The PCI values of the novice group were higher than the PCI values of the experienced group when answering arithmetic questions ($p = 0.005$). There was no significant difference between the intermediate and experienced groups ($p = 0.443$). In addition, there was no significant difference between the intermediate group and the novice group ($p = 0.128$).

Medical questions The PCI values of novices were significantly higher than of experienced participants for the medical questions ($p = 0.003$). This same pattern was observed when comparing the PCI results of novices to intermediate participants ($p = 0.023$). There was no significant difference when the PCI values of intermediate participants were compared to experienced participants ($p = 0.851$).

Variation across item difficulty, correctness and question type

Overall, difficult questions were associated with a substantially higher mean PCI than easy questions (123.40 vs. 45.14; $p < 0.001$). Similarly, incorrectly answered questions were associated with a higher PCI compared to correctly answered questions (152.51 vs. 66.04; $p < 0.001$). There was no significant difference in PCI when the arithmetic questions were compared to the medical questions (101.52 vs. 89.69; $p = 0.42$). See Table 3 for details.

Internal structure evidence

Of 232 possible data points, 10 were missing in the PCI data set because of poor pupillary size output quality in the raw data. Values were not normally distributed and were skewed toward smaller PCIs. There were a limited number of outliers, all on the high PCI side. See "Appendix 3" for additional details.

Table 3 Effect of question difficulty, correctness and type on mean pupillary change index and mean Paas score for all subgroups

	PCI mean (SD) in % seconds	Paas score mean (SD)
Difficult questions	123.40 (108.26)	5.40 (1.80)
Easy questions	45.14 (37.52)	3.62 (1.62)
Easy versus difficult	$p < 0.001$	$p < 0.001$
Correct responses	66.04 (69.46)	4.22 (1.80)
Incorrect responses	152.51 (117.82)	5.89 (1.74)
Correct versus incorrect	$p < 0.001$	$p < 0.001$
Arithmetic questions	101.52 (111.22)	4.89 (2.32)
Medical questions	89.69 (89.61)	4.66 (1.80)
Arithmetic versus medical	$p = 0.42$	$p = 0.43$

PCI pupillary change index, *SD* standard deviation

Additional evidence for Paas scale

Variation across training subgroup and question type

Arithmetic questions Similar to the PCI findings, the Paas scale results for novices were higher than for experienced participants when answering arithmetic questions ($p = 0.040$). There was also no significant difference between the intermediate and experienced groups ($p = 0.549$). In addition, there was no significant difference between the intermediate group and the novice group ($p = 0.365$).

Medical questions In keeping with the PCI results, the Paas scale values of novices were significantly higher than the Paas scale values of experienced participants for the medical questions ($p < 0.001$). This same pattern was observed when comparing the Paas scale results of novices to intermediate participants ($p = 0.006$). There was no significant difference when the Paas scale values of intermediate participants were compared to experienced participants ($p = 0.279$).

Variation across item difficulty, correctness and question type

In keeping with the PCI results, difficult questions were associated with a higher Paas scale rating than easy questions (5.40 vs. 3.62; $p < 0.001$). Similarly, incorrectly answered questions were associated with a higher Paas scale rating than correctly answered questions (5.89 vs. 4.22; $p < 0.001$). There was no significant difference in Paas scale rating when the arithmetic questions were compared to the medical questions (4.89 vs. 4.66; $p = 0.43$). See Table 3 for details.

Internal structure evidence

All 232 possible data points for the Paas scale were collected, with no missing values. The data were normally distributed, with no outliers. See "Appendix 3" for additional details.

Peak pupil size analysis

Analysis of peak pupil size data for the arithmetic questions revealed no significant differences between any of the subgroups (novice and intermediate $p = 0.15$; novice and experienced $p = 0.08$; intermediate and experienced $p = 0.97$).

Analysis of peak pupil size for the medical questions revealed a significant difference in peak pupil size between the novice group and the experienced group ($p = 0.002$). There was no significant difference between peak pupil size in the novice and intermediate groups ($p = 0.38$) or the intermediate and experienced groups ($p = 0.10$). See Table 4 for descriptive statistics.

Discussion

In this study, we compared two methods of cognitive load quantification—a physiologic measure (pupillometry with time to response) and a psychometric measure (Paas scale). In addition, we examined the relationship between cognitive load measures and experience

Table 4 Pupillary change index, Paas score and peak pupillary size by subgroup and question type

	PCI mean (SD) in % seconds	Paas score mean (SD)	Peak pupillary size mean in % (SD)
Novice (arithmetic)	160.70 (152.10)	5.84 (2.31)	112.55 (SD = 14.03)
Intermediate (arithmetic)	93.60 (80.02)	4.82 (2.13)	106.66 (SD = 16.56)
Experienced (arithmetic)	52.44 (52.09)	4.05 (2.24)	105.91 (SD = 8.09)
Novice (medical)	121.39 (113.32)	5.46 (1.75)	101.76 (10.89)
Intermediate (medical)	76.88 (67.60)	4.47 (1.83)	100.05 (6.69)
Experienced (medical)	67.73 (67.93)	3.98 (1.50)	97.33 (10.46)

PCI pupillary change index, *SD* standard deviation

levels of the participants. The goal was to add to the body of literature that supports the validity of using these techniques to measure cognitive load, using Cook's review of the Messick framework as a guide.

A direct comparison of the PCI and Paas scale values revealed a strong positive correlation. Further analysis revealed that this correlation was consistent within subgroups of experience level. This suggests that both the PCI and the Paas scale measured the same construct to some degree. Given previous work in this field as well as the confirmatory results from this experiment, it is likely that this construct is indeed overall cognitive load, which is thought to represent the sum of intrinsic and extraneous cognitive load (Sweller 2010). We suggest that the correlation found between the two instruments studied provides evidence for validity with respect to Cook's description of "relations to other variables" as both variables are commonly used in the literature to quantify cognitive load.

We also found that difficult arithmetic and difficult medical questions resulted in increased cognitive load compared to easier questions, both when analyzed using the PCI as well as the Paas questionnaire. Using both pupillometry and psychometric analysis, we found that incorrectly answered questions caused participants to experience more cognitive load than questions they answered correctly. Finally, no difference was found when questions were divided into arithmetic and medical subgroups. Pupillometry and psychometric results followed the same patterns for all of these analyses. These findings are reassuring as they are in keeping with results from previous studies that show similar trends (Szulewski et al. 2015). In sum, difficult and incorrectly answered questions caused participants to experience greater cognitive load, regardless of question type. These results are expected and bolster the validity argument from a "relations to other variables" source.

The PCI and Paas scale data for the arithmetic questions showed that novices experienced higher cognitive load compared to experts when answering questions despite no significant difference in performance between the groups. Of note, there was no significant difference between the peak pupil size between the groups when answering the arithmetic questions. The lack of significant difference in performance on the arithmetic questions makes sense as participants were divided into groups based on resuscitation medicine experience, not arithmetic experience. The lower cognitive load experienced by the experienced group compared to the novice group for the arithmetic questions may be related to age. It is possible that these older participants were either better at arithmetic or

that physiologic changes of ageing were responsible for a less responsive pupil. The lack of difference in the peak pupil size analysis for the arithmetic questions makes this latter point less likely. Importantly, both PCI and Paas scale data were consistent with one another when considering the arithmetic questions.

An analysis of the PCI and Paas scale data for the medical questions suggested that novices had higher cognitive load than both the intermediate and experienced groups. Further, no significant difference was found in cognitive load of the intermediate group compared to the expert group for these medical questions. Analogous to these trends, novices performed significantly worse than both the experienced and intermediate groups. Conceptually, these findings may be explained by the fact that novices have relative inexperience with resuscitation medicine content material, leading them both to perform more poorly and to experience higher cognitive load when attempting to find solutions to problems. The lack of difference in cognitive load of medical questions on physicians in the intermediate versus the experienced group is not surprising given that both these groups are well versed in the content material comprised in the experimental test. The intermediate group's significantly increased score over the experienced group is likely related to this group's relative proximity to their specialty examinations. Finally, the medical question analysis revealed that novices experienced a significantly increased peak pupil size compared to the experienced group. This difference was not present when answering arithmetic questions.

Together, these observations emphasize that cognitive load measurement by both physiologic and psychometric tools acts in a way that is expected and explainable across groups of physicians with varying levels of experience as well as between question types (domain-specific medical questions vs. arithmetic questions). These patterns provide some evidence of what Cook would call validity from a "relations to other variables" source. That is, the observed patterns in the data varied across groups of participants with different training status as well as with question type, as expected.

The data distribution presented in "Appendix 3" provides some validity evidence from an *internal structure* source. The analysis captured all Paas responses and missed less than 5 % of PCI values (due to poor data quality). No outliers were identified in the Paas responses and a relatively small number of outliers were identified in the PCI analysis.

Finally, this study provides some indirect evidence of *response process* as a source of validity. Response process, as a source of validity, exists when the actions and thought processes of participants align with the intended measured construct. In this study, though participants were not asked to specifically describe their thought processes (as may have been done with a think-aloud protocol), the Paas survey asked participants to rate investment of cognitive resources, which is what the pupillometry metrics were designed to measure.

Although this paper provides some evidence to support that both physiologic and psychometric measures of cognitive load quantify cognitive load as a construct to some degree, each has its own strengths and limitations. Psychometric scales are easy to use and cheap to implement, whereas pupillometry is expensive and not practical for routine use in the real world (although this will likely change as the cost of the technology decreases). On the other hand, psychometric scales are prone to participant manipulation and only provide a single cognitive load measurement. Conversely, pupillometry has the advantage of being objective, difficult to manipulate and provides real-time data throughout the peaks and troughs of cognitive effort during the completion of a task.

Importantly, we found the pupillometric data analysis to be time-consuming and difficult to automate with programmed code in this experiment as a result of changing baseline pupil sizes between questions. Changes to the experimental design by minimizing interruptions and more consistently defining pupil baseline (possibly with a standardized cognitive task), may facilitate data extraction in the future. Until this is further addressed, it would be a reasonable choice to use the Paas questionnaire as a means to determine cognitive load in a test-taking setting, especially if a general understanding (as opposed to a real-time and detailed) assessment is sought. The peak pupil size variable, which is much easier to extract, is another available option if a physiologic variable is desired. Ultimately, a real-time physiologic measure obtained unobtrusively, like pupillometry, has powerful implications for the delivery and assessment of learning within CLT.

Our study has certain limitations. To begin, we were unable to control for participant age between groups. This is unavoidable given the inherent nature of experience, but it does raise questions about the possible confounding effects of age on both pupillometry analysis as well as questionnaire responses. Although we recognize that there might be an effect, we believe it to be small given that pupillary size and psychometric responses varied in the same direction throughout this study. Further, the fact that there was no statistical difference in peak pupil size in the arithmetic questions between groups (but there was for the medical questions) raises doubts that the PCI findings are solely caused by physiological pupillary changes related to ageing.

Secondly, we used known-groups comparisons in part of our analysis of validity from a "relations to other variables" source. Though necessary, this type of comparison (on its own) is known to be non-specific and inconclusive because of possible confounding effects (Cook 2015). The analysis of training-relevant (medical) and training-irrelevant (arithmetic) questions strengthens this argument, but the possibility of confounding remains.

In addition, we did not distinguish between the types of cognitive load (intrinsic vs. extraneous vs. germane) and instead chose to focus on total measurable cognitive load. Although this strategy was necessary in order to investigate the research question, a deeper analysis using an instrument designed to differentiate types of cognitive load like the one proposed by Leppink et al. (2014) may have been beneficial—especially given findings in recent literature that have brought into question whether psychometric data may actually be measuring intrinsic, as opposed to total, cognitive load in certain settings (Naismith et al. 2015).

The fact that two potential participants from the experienced group and one potential participant from the intermediate group declined to participate in the study could lead to selection bias. In addition, the low proportion of female participants in this study has the potential to skew the results; however, we do not know of any literature that supports a difference in pupillary responses based on sex.

Finally, although we found a strong positive relationship between our two measures and thus were able to conclude that both sets of data are likely measuring the same construct to a large degree, we cannot absolutely confirm that this construct is indeed overall cognitive load, as opposed to another, related concept. Despite this, we feel that our data triangulates the available evidence and strengthens the argument that overall cognitive load is, in large part, the construct that is being measured by these techniques. From the perspective of Cook's review of Messick's framework of validity, although we provide some evidence of *internal structure*, *response process* and *relations to other*

variables, this study does not address *content* and *consequences* as sources for validity evidence.

Our work focused on cognitive load measurement in a tightly controlled (not true-to-clinical-life) MCQ environment. Future studies that measure cognitive load in real (or at least simulated) medical emergencies could provide more accurate insights into medical decision-making and in situ cognitive load. In addition, pupillometry has the potential to complement research in visual expertise (Gegenfurtner et al. in press). It is well recognized that novices and experts have different visual gaze patterns in a variety of professional domains (Gegenfurtner et al. 2011, 2013; Gegenfurtner and Szulewski 2016; Kok et al. 2012). Measuring cognitive load via pupillometry while analyzing participants' visual patterns as they perform selected tasks could provide further insights into the cognitive process and how it changes with expertise. This type of experiment would be fairly easy to perform as the eye-tracking tool used in this study can track gaze behaviours and record pupillometry data simultaneously.

The currently accepted understanding of validity places an emphasis on construct validity as the "whole" of the validity argument (Downing 2003). Based on the results of this study, we have been able to provide further evidence of construct validity that the PCI and the psychometric Paas scale are indeed reasonable surrogate markers of cognitive load. Researchers and educators can have increased confidence using either measure depending on the context and the purpose of their study, as they both appear to measure the same construct.

Conclusion

Comparing the measurement of a construct thought to represent cognitive load on medical professionals (both with a pupillometry-based physiologic tool as well as a psychometric survey) reveals a strong positive correlation between the two techniques as well as expected patterns based on question type, difficulty, correctness of answers and training status. This provides evidence that the construct being measured in both cases is related. Overall, the results support the validity of using data obtained using either technique as a surrogate for cognitive load. Further study into the subtypes of cognitive load in medical testing environments as well as cognitive load measurement in real-life clinical scenarios has the potential to provide new insights into the clinical decision-making process.

Acknowledgments The authors would like to thank Wilma Hopman for assistance with statistical analysis, Bence Linder for development and implementation of the algorithm to smooth the raw pupillometry data and to calculate peak pupillary size, as well as Dr. Jimmie Leppink for advice about experimental design. The authors would also like to acknowledge the Kingston Resuscitation Institute for providing access to the pupillometry device and research assistants.

Appendix 1

Psychometric survey used in the study, adapted from Paas (1992).

Please choose the category (1, 2, 3, 4, 5, 6, 7, 8, or 9) that applies to you:

In the exercise that just finished, I invested:

1. very, very low mental effort

2. very low mental effort

3. low mental effort

4. rather low mental effort

5. neither low nor high mental effort

6. rather high mental effort

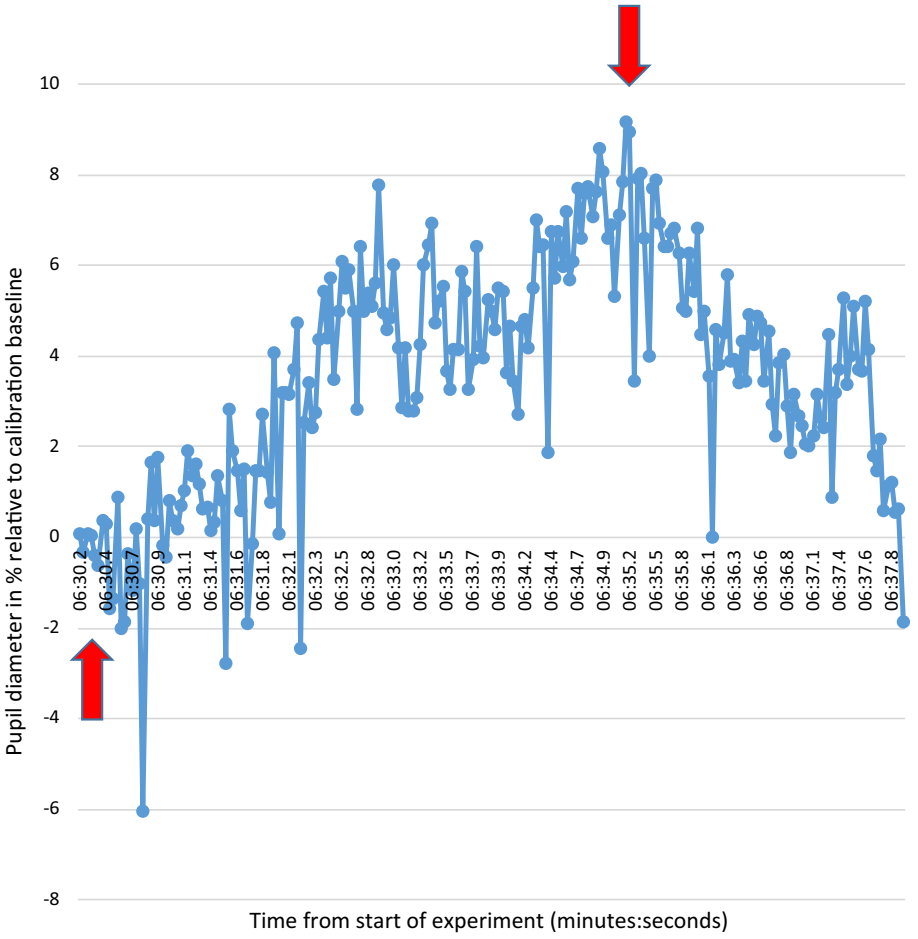
7. high mental effort

8. very high mental effort

9. very, very high mental effort

Appendix 2

Example of raw pupillometry data obtained from one experienced participant for one medical question. The first arrow represents the time the question appeared on the screen. The second arrow represents the point at which the participant verbalized his answer. As the participant experiences increasing cognitive load during the thought process, the pupil diameter increases in size. When the participant verbalizes the answer to the question, pupil size decreases again. The quantitative cognitive load measurement used in the pupillometry arm of this study can be conceptualized as the area under the curve between these two arrows (referred to as pupillary change index in this manuscript).

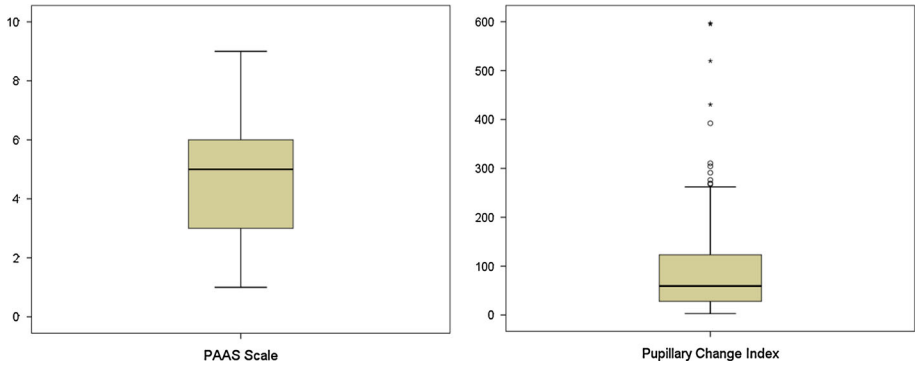


Appendix 3

Data distribution for Paas and Pupillary Change Index scales:

	Paas	PCI
Valid data points	232	222
Missing data points	0	10
Minimum value	1	2.9
Maximum value	9	597.2
Mean	4.7	92.7
Standard deviation	1.9	95.5
25th percentile	3	27.7
Median	5	59.4
75th percentile	6	124.2

Boxplots of the data distribution of the Paas and Pupillary Change Index scales showing outliers:



References

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*(5), 389–400.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53–61.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L: Erlbaum.

Cook, D. A. (2015). Much ado about differences: Why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education, 20*(3), 829–834.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, *119*(2), 166.e7–166.e16.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38*(2), 105–134.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830–837.

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

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*(7/8), 114.

Gegenfurtner, A., Kok, E., Van Geel, K., De Bruin, A., Jarodzka, H., Szulewski, A., & Van Merriënboer, J. J. G. (in press). The challenges of studying visual expertise in medical image diagnosis. *Medical Education*.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523–552.

Gegenfurtner, A., & Seppänen, M. (2013). Transfer of expertise: An eye tracking and think aloud study using dynamic medical visualizations. *Computers and Education, 63*, 393–403.

Gegenfurtner, A., Siewiorek, A., Lehtinen, E., & Säljö, R. (2013). Assessing the quality of expertise differences in the comprehension of medical visualizations. *Vocations and Learning, 6*(1), 37–54.

Gegenfurtner, A., & Szulewski, A. (2016). Visual expertise and the Quiet Eye in sports – comment on Vickers. *Current Issues in Sport Science*, *1*, 108. doi:10.15203/CISS_2016.108.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, *212*, 46–54.

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

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585.

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 and 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.

Kok, E. M., Bruin, A. B., Robben, S. G., & Merriënboer, J. J. (2012). Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. *Applied Cognitive Psychology, 26*(6), 854–862.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18–27.

Laxmisan, A., Hakimzada, F., Sayan, O. R., Green, R. A., Zhang, J., & Patel, V. L. (2007). The multitasking clinician: Decision-making and cognitive demand during and after team handoffs in emergency care. *International Journal of Medical Informatics, 76*(11), 801–811.

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.

Naismith, L. M., & Cavalcanti, R. B. (2015). Validity of cognitive load measures in simulation-based training: A systematic review. *Academic Medicine, 90*(11), S24–S35.

Naismith, L. M., Cheung, J. J., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education, 49*(8), 805–814.

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*(4), 418–427.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429.

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.

Perry, S. J., Wears, R. L., Croskerry, P., & Shapiro, M. J. (2013). Process Improvement and Patient Safety. In J. Marx, R. Walls & R. Hockberger (Eds.), *Rosen's Emergency Medicine - Concepts and Clinical Practice* (8 Edn., Vol. 2, pp. 2505–2511). Philadelphia: Elsevier Health Sciences.

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A., & Pappas, J. (2013). Characterizing novice-expert differences in macrocognition: An exploratory study of cognitive work in the emergency department. *Annals of Emergency Medicine*, *61*(1), 96–109.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123–138.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

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*, *2*(1), 8–11.

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.

Tuovinen, J., & Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science, 32*(1–2), 133–152. doi:10.1023/B:TRUC.0000021813.24669.62.

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.