LOOKING BEYOND SELF-REPORTED COGNITIVE LOAD: INVESTIGATING THE USE OF EYE TRACKING IN THE STUDY OF DESIGN REPRESENTATIONS IN ENGINEERING DESIGN

Cass, Madison (1);
Prabhu, Rohan (2)

1: Neuroscience Program, Lafayette College;
2: Department of Mechanical Engineering, Lafayette College

ABSTRACT

Designers are experiencing greater mental demands given the complexity of design tools, necessitating the study of cognitive load in design. Researchers have identified task- and designer-related factors that affect cognitive load; however, these studies primarily use self-reported measures that could be inaccurate and incomplete. Little research has tested the accuracy and completeness of self-reported measures and we aim to explore this gap. Towards this aim, we seek to answer the question: How does cognitive load vary based on the different design representations used, and do these differences depend on the measure of cognitive load? From our results, we see that the design representations vary in the range of cognitive load experienced by designers when using them. Moreover, this role of the range of cognitive load variance was observed given our use of pupil diameter. These findings call for the use of a multi-modal approach for measuring cognitive load with the combined use of subjective (e.g., self-report) and objective measures (e.g., physiological measures), as well as the use of both retrospective (e.g., self-report) and concurrent measures (e.g., physiological measures).

Keywords: Design for Additive Manufacturing (DfAM), Design cognition, Industry 4.0, cognitive load, design representation

Contact:
Prabhu, Rohan
Lafayette College
United States of America
prabhur@lafayette.edu

Cite this article: Cass, M., Prabhu, R. (2023) 'Looking beyond Self-Reported Cognitive Load: Investigating the Use of Eye Tracking in the Study of Design Representations in Engineering Design', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.248

1 INTRODUCTION

The field of engineering design is becoming increasingly complex through the emergence of new technological developments, such as additive manufacturing (Simpson et al., 2016) and machine learning (Panchal et al., 2019). These advancements largely benefit designers as they help designers execute complex tasks in relatively simple ways (Gann et al., 2004) and enhance the effectiveness of the design process (Gardan, 2016). However, these technologies pose new challenges to designers due to phenomena such as choice overload (Erhan et al., 2017) and innovation overload (Herbig and Kramer, 1992). That is, designers must now sift through large quantities of information and data to make design decisions. These challenges could, in turn, contribute to cognitive overload, wherein designers reach their neural processing limits and can no longer effectively perform the task at hand (Calpin and Menold, 2022). Given these greater mental demands, there has emerged a need to study designers' cognitive load in light of these advanced design and manufacturing technologies.

Motivated by this need, researchers have studied cognitive load across different stages of the design process. For example, Calpin and Menold (2022) find that the early stages of the design process (e.g., concept generation) could require higher levels of cognitive load compared to later design stages (e.g., prototyping). Moreover, they posit that while differences in cognitive load could impact conceptual design performance, this impact is not as significant on prototyping performance. Therefore, the different stages in the design process could impose different levels of cognitive load on designers and one possible source of these differences could be the design representations used in these stages.

Design representations play an important role in engineering design as they provide designers with a medium to translate their domain knowledge and skills into solutions (McKoy et al., 2001). Prior work suggests that design representations vary in the affordances they provide to designers. For example, while higher-fidelity representations tend to be used more in geometry- and manufacturing-focused decision-making, low-fidelity representations could suffice in conceptual decision-making (Hannah et al., 2012). Additionally, design representations have also been shown to require different and sometimes non-interchangeable cognitive skills such as spatial visualization (Sorby, 1999).

Motivated by the differences in the affordances offered by different design representations and the cognitive skills required in employing them, researchers have also studied the impact of using these design representations on designers' cognitive load. For example, in a within-subjects study, Bilda and Gero (2007) find that during a design task, designers experienced higher cognitive load when they were not allowed to sketch compared to when they were allowed to sketch. In a similar study, Nolte and McComb (2021) find that generating a concept and building a physical model required the highest amount of cognitive load. Contrary to these findings, Mohamed-Ahmed et al. (2013) observe that CAD, sketching, and prototyping required similar cognitive loads. These findings further reinforce the notion that design representations could play an important role in determining the cognitive load imposed in different stages of engineering design.

Collectively, this research on cognitive load in design has emphasized the importance of measuring designers' cognitive load, particularly to identify the design tasks that impose the greatest levels of cognitive demands. However, the *range* in cognitive load within and across different design tasks remains little explored. Additionally, an overwhelming number of these design studies use self-report measures, which might not be sufficient to quantify cognitive load (de Waard and Lewis-Evans, 2014). This over-emphasis on self-reported measures has resulted in alternative methods, such as physiological markers (e.g., pupil diameter and heart rate), being seldom used despite their efficacy in providing an objective and concurrent measurement. Furthermore, given this scarce use of physiological measures, their role and utility in design cognition research remain little explored.

Our aim in this paper is to explore this research gap by answering the research question **(RQ): How does cognitive load vary based on the different design representations used, and do these differences depend on the measure of cognitive load?** We hypothesize that the NASA-TLX global scores and pupil diameter would vary across the four design representations (i.e., sketching, writing,

https://doi.org/10.1017/pds.2023.248 Published online by Cambridge University Press

CAD modelling, and build simulation). This hypothesis is based on prior work where some design representations were found to be more cognitively demanding than others (Nolte and McComb, 2021). However, the directionality of this relationship is uncertain since other research (e.g., by Mohamed-Ahmed et al. (2013)) has observed no differences in cognitive load across design representations.

2 EXPERIMENTAL METHODS

To answer our RQ, we conducted an experimental study at a small private liberal arts college in the north-eastern United States of America. The experimental protocol was reviewed and approved by the Institutional Review Board before it was conducted, and its details are discussed next.

2.1 Participants

Fourth-year mechanical engineering students (N=10) were recruited to participate in this study. When asked about their gender identity, 20% of participants self-identified as female, 70% as male, and 10% as non-binary. Only fourth-year students were included in the study to ensure that they all had prior exposure to CAD (Fusion 360™) through a second-year course on design and manufacturing. Participation was voluntary and each participant received a $20 gift card as compensation.

2.2 Procedure

The study was conducted in the spring semester of 2022 and comprised two stages: (1) a DfAM task, and (2) a post-intervention survey. Informed consent was obtained from the participants before the experiment was conducted. The details of each stage of the experiment are discussed next.

2.2.1 Design for additive manufacturing task

Participants were individually tasked with completing a 60-minute DfAM task. We chose to perform an open-ended design task as opposed to a fully controlled experiment motivated by the need to study design in different naturalistic contexts and environments, as suggested by Reich (2022). Specifically, we provided participants with the following design prompt:

> *Design a fully 3D-printable solution to enable hands-free viewing of content on a smartphone. You can design your solution to fit any phone of your choice. Your design should use the least amount of print material possible and print as fast as possible.*

The design task was similar to the open-ended design task proposed by Prabhu et al. (2020b), with few explicit manufacturing constraints and functional requirements. This design task was chosen as it does not require domain expertise as everyone is familiar with the problem context of a cell phone holder. Additionally, this task did not have many restrictions, which allowed participants to use a wide range of principles to solve the stated problem. Despite being open-ended, we prescribed some constraints in the task to provide some guidance to the participants' concept generation. This decision was informed by the findings of Onarheim (2012), who suggest that a combination of inflexible and flexible constraints fuels creativity by providing direction to the design process without restricting creative freedom.

During the design task, participants wore the Tobii Pro Glasses 3 eye tracker which recorded data every millisecond. The calibration and recording procedure outlined by Tobii Pro was followed to ensure that accurate and precise data were collected. Since the eye tracker can only be used on participants who do not wear corrective glasses, we attempted to recruit participants who could perform the task without prescription glasses. However, when recruiting, participants were given the option to wear contact lenses. Since this experiment occurred during the COVID pandemic, most participants wore their face masks; however, participants were given the freedom to remove their masks if they felt comfortable (e.g., to avoid fogging the glasses).

During the design task, participants had access to a ruler, a protractor, a calculator, plain paper, and writing instruments. No music or internet searching was allowed unless it involved searching for basic information such as dimensions and unit conversions. As outlined in the design task, participants were expected to create a fully completed CAD model and a build file for the solution by the end of the

https://doi.org/10.1017/pds.2023.248 Published online by Cambridge University Press

task. In the case that participants were not familiar with creating a build file, they were provided a how-to guide on PrusaSlicer, a 3D build simulation software (guide available here: sites.lafayette.edu/kidd-lab).

2.2.2 Post-intervention survey

After completing the design task, participants were asked to complete a post-intervention survey. This survey consisted of four NASA-TLX (Hart and Staveland, 1988) assessments to evaluate the participants' perceived cognitive load when working with the four possible design representations: (1) sketching, (2) writing notes, (3) designing a CAD model, and (4) simulating an AM build. Participants were asked to report cognitive load for each stage on six sub-components: (1) mental demand, (2) physical demand, (3) temporal demand, (4) performance, (5) effort, and (6) frustration. Participants were asked to report their perceived cognitive load on each sub-component on a 20-point scale with 1 = "low" to 20 = "high" (see Section 2.3.1 for details about the NASA-TLX and its sub-components).

2.3 Metrics and coding scheme

The data collected was evaluated using the metrics discussed next.

2.3.1 Self-reported cognitive load

We used the NASA-TLX questionnaire as our self-report measure for cognitive load. The NASA-TLX uses a multi-dimensional approach to measure cognitive load, making it a preferred method of mental workload assessment in a vast array of contexts (Hart, 2006). We chose to use the NASA-TLX because of its high reliability, sensitivity to different levels of mental workload, and ease of use (Devos et al., 2020; Ikuma et al., 2009; Rubio et al., 2004; Wierwille and Eggemeier, 1993). As outlined in Table 1, the NASA-TLX is composed of six sub-components, each of which contributes information independent of the other sub-components (Hart and Staveland, 1988). This unique structure of the NASA-TLX provides a clear advantage over its counterpart questionnaires, such as the SWAT (Reid and Nygren, 1988) which do not give sub-component level scores (Rubio et al., 2004).

We made comparisons at the global level of workload using an unweighted score because Hart and Staveland (1988) found that the weighted score does not greatly improve the statistical sensitivity of the experimental variables. Additionally, the unweighted workload score has been found to be more reliable and sensitive than the weighted workload score (Devos et al., 2020; Ikuma et al., 2009).

Table 1. Definitions of the six sub-components from the NASA task load index

Sub-Component	Description
Mental Demand	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	How much physical activity was required? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

2.3.2 Eye tracking as a measure of cognitive load

In addition to self-reported cognitive load, we used eye tracking as a source of physiological measurement of cognitive load. Eye tracking is a reliable and non-intrusive measure of mental workload that enables the remote collection of several eye movement parameters, such as pupil

https://doi.org/10.1017/pds.2023.248 Published online by Cambridge University Press

dilation and frequency of fixations (Cowley et al., 2016). Of these measures, we used pupil diameter given the significant evidence of its positive relationship to cognitive load (Beatty, 1982; Beatty and Lucero-Wagoner, 2000; Cabestrero *et al.*, 2009; Hess and Polt, 1964; Kahneman, 1973).

Once collected, the data were processed using Tobii Pro Lab. We used the fixation gaze filter that groups raw gaze points into fixations based on whether the eyes remain focused on the same location. The use of the fixation gaze filter results in an average of two to three fixations per second. We chose the fixation gaze filter because it is more appropriate for situations involving still movements (e.g., looking at a computer screen), compared to the attention gaze filter (Tobii Pro AB, 2022).

From the collected recordings, we exported the measured pupil diameter of each eye per fixation. Next, we performed factor analysis on the left and right pupil diameter to reduce dimensionality and identify one latent factor to represent cognitive load. We chose factor analysis as it can effectively uncover underlying factors in the data while still retaining possible variance in the original variables (Alavi et al., 2020). Before performing factor analysis, we tested the assumptions of the analysis. Specifically, the sampling adequacy of the data was assessed using a Kaiser-Meyer-Olkin (KMO) measure > 0.50 (Kaiser, 1974), and we determined that the two variables were correlated for every participant using Bartlett's test of sphericity ($p < 0.05$; Tobias and Carlson, 1969). Since we performed factor analysis on two variables, only one latent factor was suggested (eigenvalues > 1). This one-factor model explained a cumulative 97.4% of the variance and was accordingly labelled 'Latent Pupil Diameter'. The factor analysis was performed independently for each participant. The internal consistency and reliability of the factor were assessed through Cronbach's $\alpha > 0.90$ (Cronbach, 1951) across all participants. After performing factor analysis, we calculated four descriptives of the latent pupil diameter: (1) mean, (2) maximum, (3) minimum, and (4) variance (maximum-minimum). These descriptives were calculated for the time each participant spent using each design representation (i.e., sketching, writing, CAD modelling, and build simulation). These descriptives of the latent pupil diameter were used as representative metrics for cognitive load measured using eye tracking.

2.3.3 Coding scheme for designers' use of design representations

In addition to the two methods of cognitive load assessment, we aimed to determine how cognitive load varies based on the design representation used. Therefore, we qualitatively coded the four design representations for each fixation using Tobii Pro Lab. We created four 'Areas of Interest,' corresponding to a design representation: (1) sketching, (2) writing, (3) CAD modelling, and (4) build simulation. We subsequently coded each fixation point according to the specific area of interest observed in the dataset. Initially, two reviewers (two undergraduate students) independently coded 10% of the eye-tracking data using the codebook presented in Table 2. Upon observing acceptable inter-rater reliability (Cohen's Kappa = 0.94; Hallgren, 2012), the primary reviewer coded the remaining data.

3 DATA ANALYSIS AND RESULTS

Our aim in this paper is to answer the RQ: *How does cognitive load vary based on the different design representations used and do these differences depend on the measure of cognitive load?* We hypothesized that there would be differences in cognitive load—both, self-reported and eye-tracked—across the four design representations (i.e., sketching, writing, CAD Modelling, and build simulation). To test this hypothesis, the data collected from the experiment were analysed using quantitative methods. Parametric tests were used, and we performed the analyses using a statistical significance level of $\alpha = 0.05$ and a 95% confidence interval. Data from all 10 participants were used in the final analysis and no participants were excluded from the analyses.

Specifically, we computed a series of mixed linear model regressions (MLMR). In each regression model, the four latent pupil diameter descriptives (i.e., mean, maximum, minimum, and variance) and the NASA-TLX global scores were taken as the dependent variables, and the design representation (i.e., sketching, writing, CAD modelling, build simulation) was used as the independent variable. We chose the MLMR because it is suitable for repeated measures analyses with missing data and small sample sizes (Magezi, 2015). From the results of the five regressions summarized in Table 3 and

https://doi.org/10.1017/pds.2023.248 Published online by Cambridge University Press

Figure 1, we observed that the maximum, minimum, and variance of latent pupil diameter varied significantly based on the design representation being used. However, no significance was observed for the mean of latent pupil diameter or NASA-TLX global score.

We performed pairwise comparisons on the variables that demonstrated significance to test differences in latent pupil diameter and NASA-TLX global scores between specific design representations. Before performing the post-hoc analyses, we tested the groups for their equality in variances using the F-test. From the results, we observed an unequal variance between our groups and therefore, we used Tamhane's T2 statistical method, a parametric post-hoc test for normally distributed data with unequal group variances (Tamhane, 1979). From the results of the post-hoc analyses, we see that the comparison of CAD modelling and writing for the variance of latent pupil diameter was statistically significant ($p < 0.01$). We also observed a significant difference between CAD modelling and simulating the build for the variance of latent pupil diameter ($p < 0.04$). Additionally, there was a significant difference between CAD modelling and writing for the scaled minimum of latent pupil diameter ($p < 0.05$). No other statistically significant pairwise comparisons were observed. The implications of these results are discussed in the next section.

Table 2. Codebook of the four design representations with definitions and examples

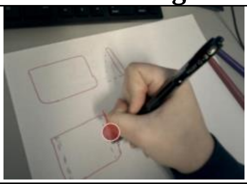
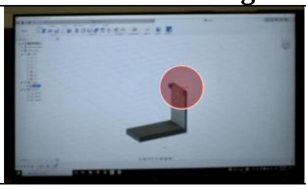
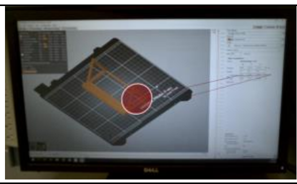
Sketching	**Writing Notes**	**CAD Modelling**	**Build Simulation**
			
Any activity in which the participant: - drew a shape - wrote dimensions of a drawn shape, or - measured their phone in between periods of sketching.	Any activity in which the participant: - underlined or wrote letters, numbers, and symbols (except dimensions).	Any activity in which the participant: - looked at the computer screen with Fusion 360™, - measured their phone when designing a CAD model, or - looked at the keyboard or their sketches/notes when designing a CAD model.	Any activity in which the participant: - looked at the computer screen with PrusaSlicer, - looked at the printed PrusaSlicer guide, - looked at the keyboard or their sketches/notes when simulating their AM build, or - looked at the screen with Fusion 360™ when simulating their build (e.g., saving the CAD model as an STL file).

Table 3. Statistical summary of the mixed linear model regression analysis

Latent Pupil Diameter	**Mean**	**Standard Error**	**Fisher's z**	**p**
Mean	-0.05	0.20	-0.23	0.82
Maximum	-0.41	0.18	-2.23	0.03
Minimum	0.44	0.19	2.29	0.02
Variance (min - min)	-0.86	0.20	-4.42	0.00
NASA-TLX Global Score	-0.46	0.31	-1.48	0.14

https://doi.org/10.1017/pds.2023.248 Published online by Cambridge University Press

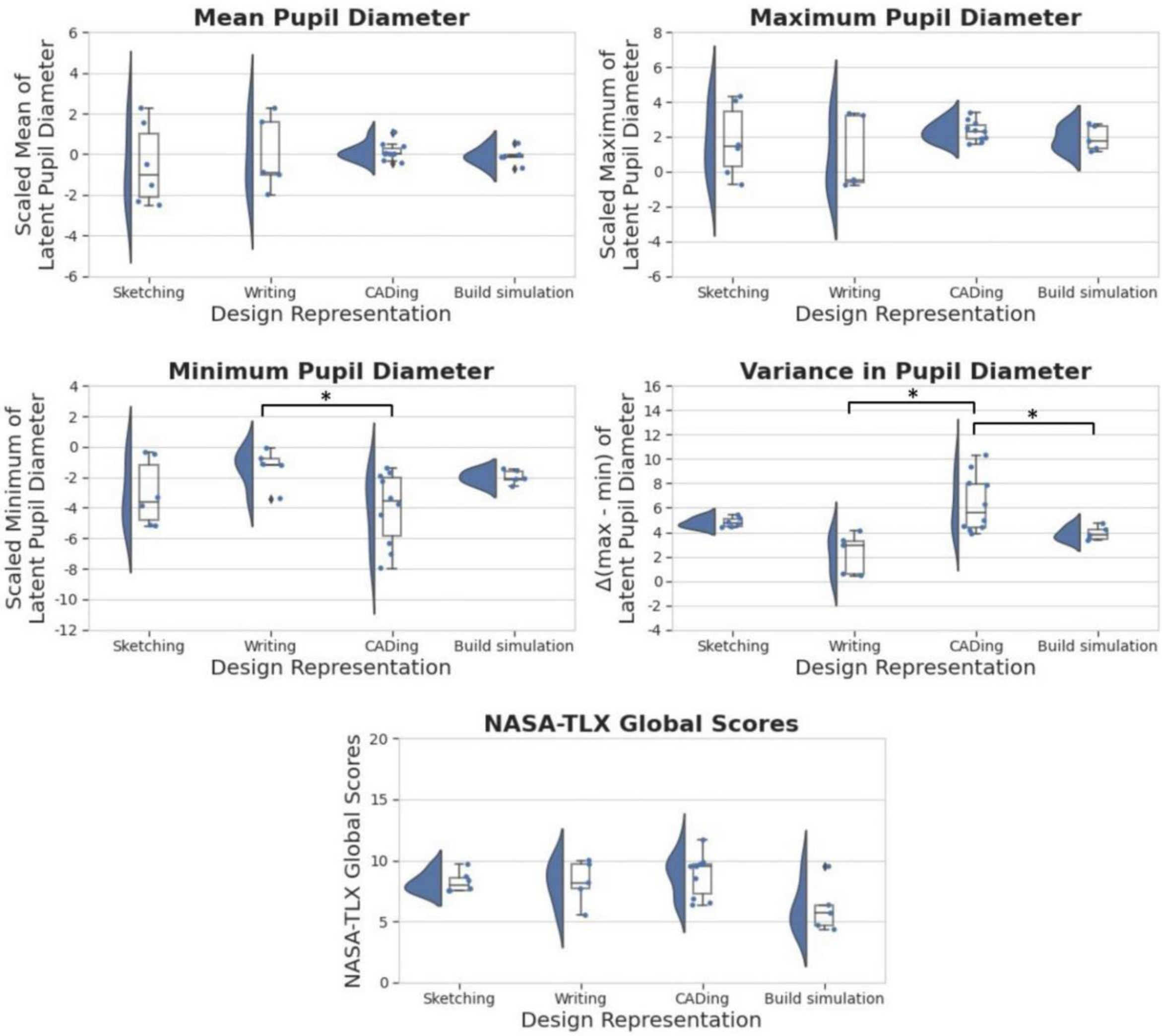


Figure 1. Comparing the descriptives of latent pupil diameter and NASA-TLX global scores across the four design representations (indicates $p < 0.05$)*

4 DISCUSSIONS AND IMPLICATIONS OF RESULTS

Our aim in this paper is to investigate differences in cognitive load when designers use different design representations and test whether these differences depend on the measure of cognitive load. The key finding from our experimental study is that design representations vary in the *range* of cognitive load required by designers when utilizing them. That is, although we did not see any significant differences in the mean, minimum, or maximum pupil diameter, we did see differences in the variance in pupil diameter (i.e., maximum - minimum).

This finding suggests the importance of variance as a descriptive statistic for pupil diameter as it can capture aspects of cognitive load that other descriptives cannot. This finding supports previous research by Hershaw and Ettenhofer (2018) who found that peak-based metrics (e.g., maximum and variance) are more reliable and sensitive compared to average-based metrics. However, few studies using pupil diameter to measure cognitive load have included variance in their analyses; instead, average-based metrics of pupil diameter are commonly used (Hershaw and Ettenhofer, 2018). Therefore, this finding presents an opportunity for the future use of variance as a metric for pupil diameter.

Additionally, we see that designers' pupil diameter showed a greater variance when they generated a CAD model of their solution compared to when they wrote notes or simulated their AM build. This finding suggests that cognitive load is experienced along a spectrum that varies based on the design representation. However, the NASA-TLX global score might not capture this cognitive load spectrum.

https://doi.org/10.1017/pds.2023.248 Published online by Cambridge University Press

Therefore, solely comparing the NASA-TLX global score across design representations is insufficient; instead, a supplemental measure that can account for variance in cognitive load, such as pupil diameter, must be used. This finding, thus, reinforces the advantages of using multiple methods to effectively measure cognitive load.

Taken together, our findings highlight the disadvantages of using a global score for NASA-TLX as a sole measure of cognitive load; accordingly, we provide three recommendations for the future study of cognitive load in design. First, we find that pupil diameter is inconsistent with the NASA-TLX global score. This inconsistency could be due to the use of a global score since it cannot adequately capture the subtleties of the different dimensions of cognitive load. Therefore, future research should explore participants' accuracies in self-evaluating their cognitive load using the individual NASA-TLX sub-component scores. Secondly, we suggest using a multi-modal approach for measuring cognitive load. To maximize the benefits of this approach, we recommend the combined use of subjective (e.g., self-report) and objective measures (e.g., physiological measures), as well as the use of both retrospective (e.g., self-report) and concurrent measures (e.g., physiological measures). Finally, we observe differences in the *range of pupil diameter* depending on the design representation under consideration. This finding suggests that designers' cognitive load could exist on a spectrum and therefore, we suggest measuring cognitive load using the variance in pupil diameter instead of a single-point measure.

5 CONCLUSIONS, LIMITATIONS, AND DIRECTIONS FOR FUTURE WORK

Our aim in this paper was to investigate differences in cognitive load when designers use different design representations and test whether these differences depend on the measure of cognitive load. From our results, we see that design representations vary in the range of cognitive load experienced by designers when utilizing them. This finding suggests the importance of variance as a metric for pupil diameter, mostly because it can capture the spectrum of cognitive load better than the other pupil diameter descriptives and the NASA-TLX global score. This finding calls for the use of the multi-modal approach for assessing cognitive load.

While this finding provides important insights into the measurement of designers' cognitive load, our study has some limitations, presenting directions for future work. First, we measured pupil diameter to understand cognitive load, but not all observed pupillary responses are solely due to changes in cognitive load. For example, prior research suggests that factors such as luminance (Cherng et al., 2020), emotional processing and attention levels (Peinkhofer et al., 2019), decision-making (Murphy et al., 2014), and metacognitive confidence (Lempert et al., 2015) could cause changes in pupil diameter. Therefore, researchers should separate the various factors of pupillary responses so that only the relevant pupillary information corresponding to cognitive load is extracted. Second, we qualitatively coded each visual fixation according to the four AOIs (i.e., design representations) using our codebook presented in Table 2. Although we coded the fixations based on our collective interpretation of the AOIs, the designers may have perceived the AOIs differently when they were completing the NASA-TLX questionnaire. This limitation would have influenced our comparisons between pupil diameter and the NASA-TLX score based on the design representation. Future research could combined the use of eye tracking with a think-aloud protocol (Gero and Milovanovic, 2020) to collect designers' perceptions.
Additionally, the AOIs only captured the design representation used without considering the purpose of employing the design representation. Prior research suggests that different design representations afford designers the ability to apply different bodies of domain knowledge (Hannah et al., 2012). Therefore, future research will test correlations between design representations and the DfAM knowledge applied by the designers using these representations. Finally, our study had a small sample of participants. This limitation is particularly exacerbated since not all the participants used each of the four design representations, therefore, providing fewer data points for analysis. Despite the small sample size, we believe that given the highly granular data obtained from eye-tracking, our findings could make an important contribution, especially when compared against suggested sample sizes suggested for interview- and field study-based research (Cash et al., 2022).

REFERENCES

Alavi, M., Visentin, D.C., Thapa, D.K., Hunt, G.E., Watson, R. and Cleary, M. (2020), "Exploratory factor analysis and principal component analysis in clinical studies: Which one should you use?", Journal of Advanced Nursing, p. jan.14377, https://dx.doi.org/10.1111/jan.14377.

Beatty, J. (1982), "Task-evoked pupillary responses, processing load, and the structure of processing resources.", Psychological Bulletin, Vol. 91 No. 2, pp. 276–292, https://dx.doi.org/10.1037/0033-2909.91.2.276.

Beatty, J. and Lucero-Wagoner, B. (2000), "The Pupillary System", Handbook of Psychophysiology, 2nd ed., Cambridge University Press, pp. 142–162.

Bilda, Z. and Gero, J.S. (2007), "The impact of working memory limitations on the design process during conceptualization", Design Studies, Vol. 28 No. 4, pp. 343–367, https://dx.doi.org/10.1016/j.destud.2007.02.005.

Cabestrero, R., Crespo, A. and Quirós, P. (2009), "Pupillary Dilation as an Index of Task Demands", Perceptual and Motor Skills, Vol. 109 No. 3, pp. 664–678, https://dx.doi.org/10.2466/pms.109.3.664-678.

Calpin, N. and Menold, J. (2022), "The Cognitive Costs of Design Tasks: The Evolution of Cognitive Load in Design and its Relationship with Design Outcomes", ASME.

Cash, P., Isaksson, O., Maier, A. and Summers, J. (2022), "Sampling in design research: Eight key considerations", Design Studies, Vol. 78, p. 101077, https://dx.doi.org/10.1016/j.destud.2021.101077.

Cherng, Y.-G., Baird, T., Chen, J.-T. and Wang, C.-A. (2020), "Background luminance effects on pupil size associated with emotion and saccade preparation", Scientific Reports, Vol. 10 No. 1, p. 15718, https://dx.doi.org/10.1038/s41598-020-72954-z.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., Barral, O., et al. (2016), "The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction", Foundations and Trends® in Human–Computer Interaction, Vol. 9 No. 3–4, pp. 151–308, https://dx.doi.org/10.1561/1100000065.

Cronbach, L.J. (1951), "Coefficient alpha and the internal structure of tests", Psychometrika, Vol. 16 No. 3, pp. 297–334, https://dx.doi.org/10.1007/BF02310555.

Devos, H., Gustafson, K., Ahmadnezhad, P., Liao, K., Mahnken, J.D., Brooks, W.M. and Burns, J.M. (2020), "Psychometric Properties of NASA-TLX and Index of Cognitive Activity as Measures of Cognitive Workload in Older Adults", Brain Sciences, Vol. 10 No. 12, p. 994, https://dx.doi.org/10.3390/brainsci10120994.

Erhan, H., Chan, J., Fung, G., Shireen, N. and Wang, I. (2017), "Understanding Cognitive Overload in Generative Design - An Epistemic Action Analysis", presented at the The Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2017: Protocols, Flows, and Glitches, Hong Kong, pp. 127–137, doi: https://doi.org/10.52842/conf.caadria.2017.127.

Gann, D., Dodgson, M. and Salter, A. (2004), "Impact of innovation technology on engineering problem solving: Lessons from high profile public projects".

Gardan, J. (2016), "Additive manufacturing technologies: state of the art and trends", International Journal of Production Research, Vol. 54 No. 10, pp. 3118–3132, https://dx.doi.org/10.1080/00207543.2015.1115909.

Gero, J.S. and Milovanovic, J. (2020), "A framework for studying design thinking through measuring designers' minds, bodies and brains", Design Science, Cambridge University Press, Vol. 6, p. e19, https://dx.doi.org/10.1017/dsj.2020.15.

Hallgren, K.A. (2012), "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial", Tutorials in Quantitative Methods for Psychology, Vol. 8 No. 1, pp. 23–34, https://dx.doi.org/10.20982/tqmp.08.1.p023.

Hannah, R., Joshi, S. and Summers, J.D. (2012), "A user study of interpretability of engineering design representations", Journal of Engineering Design, Vol. 23 No. 6, pp. 443–468, https://dx.doi.org/10.1080/09544828.2011.615302.

Hart, S.G. (2006), "Nasa-Task Load Index (NASA-TLX); 20 Years Later", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 50 No. 9, pp. 904–908, https://dx.doi.org/10.1177/154193120605000909.

Hart, S.G. and Staveland, L.E. (1988), "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research", Advances in Psychology, Vol. 52, Elsevier, pp. 139–183, https://dx.doi.org/10.1016/S0166-4115(08)62386-9.

Herbig, P.A. and Kramer, H. (1992), "The phenomenon of innovation overload", Technology in Society, Vol. 14 No. 4, pp. 441–461, https://dx.doi.org/10.1016/0160-791X(92)90038-C.

Hershaw, J.N. and Ettenhofer, M.L. (2018), "Insights into cognitive pupillometry: Evaluation of the utility of pupillary metrics for assessing cognitive load in normative and clinical samples", International Journal of Psychophysiology, Vol. 134, pp. 62–78, https://dx.doi.org/10.1016/j.ijpsycho.2018.10.008.

Hess, E.H. and Polt, J.M. (1964), "Pupil Size in Relation to Mental Activity during Simple Problem-Solving", Science, Vol. 143 No. 3611, pp. 1190–1192, https://dx.doi.org/10.1126/science.143.3611.1190.

Ikuma, L.H., Nussbaum, M.A. and Babski-Reeves, K.L. (2009), “Reliability of physiological and subjective responses to physical and psychosocial exposures during a simulated manufacturing task”, International Journal of Industrial Ergonomics, Vol. 39 No. 5, pp. 813–820, https://dx.doi.org/10.1016/j.ergon.2009.02.005.

Kahneman, D. (1973), Attention and Effort, Prentice Hall, Englewood Cliffs, NJ.

Kaiser, H.F. (1974), “An index of factorial simplicity”, Psychometrika, Vol. 39 No. 1, pp. 31–36, https://dx.doi.org/10.1007/BF02291575.

Lempert, K.M., Chen, Y.L. and Fleming, S.M. (2015), “Relating Pupil Dilation and Metacognitive Confidence during Auditory Decision-Making”, PLOS ONE, Public Library of Science, Vol. 10 No. 5, p. e0126588, https://dx.doi.org/10.1371/journal.pone.0126588.

Magezi, D.A. (2015), “Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (LMMgui)”, Frontiers in Psychology, Vol. 6.

McKoy, F.L., Vargas-Hernández, N., Summers, J.D. and Shah, J.J. (2001), “Influence of Design Representation on Effectiveness of Idea Generation”, Volume 4: 13th International Conference on Design Theory and Methodology, presented at the ASME 2001 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, Pittsburgh, Pennsylvania, USA, pp. 39–48, https://dx.doi.org/10.1115/DETC2001/DTM-21685.

Mohamed-Ahmed, A., Bonnardel, N., Côté, P. and Tremblay, S. (2013), “Cognitive load management and architectural design outcomes”, International Journal of Design Creativity and Innovation, Vol. 1 No. 3, pp. 160–176, https://dx.doi.org/10.1080/21650349.2013.797013.

Murphy, P.R., Vandekerckhove, J. and Nieuwenhuis, S. (2014), “Pupil-Linked Arousal Determines Variability in Perceptual Decision Making”, PLOS Computational Biology, Public Library of Science, Vol. 10 No. 9, p. e1003854, https://dx.doi.org/10.1371/journal.pcbi.1003854.

Nolte, H. and McComb, C. (2021), “The cognitive experience of engineering design: an examination of first-year student stress across principal activities of the engineering design process”, Design Science, Vol. 7, p. e3, https://dx.doi.org/10.1017/dsj.2020.32.

Onarheim, B. (2012), “Creativity from constraints in engineering design: lessons learned at Coloplast”, Journal of Engineering Design, Vol. 23 No. 4, pp. 323–336, https://dx.doi.org/10.1080/09544828.2011.631904.

Panchal, J.H., Fuge, M., Liu, Y., Missoum, S. and Tucker, C. (Eds.). (2019), “Special Issue: Machine Learning for Engineering Design”, Journal of Mechanical Design, Vol. 141 No. 11, https://dx.doi.org/10.1115/1.4044690.

Peinkhofer, C., Knudsen, G.M., Moretti, R. and Kondziella, D. (2019), “Cortical modulation of pupillary function: systematic review”, PeerJ, Vol. 7, p. e6882, https://dx.doi.org/10.7717/peerj.6882.

Prabhu, R., Miller, S.R., Simpson, T.W. and Meisel, N.A. (2020), “Complex Solutions for Complex Problems? Exploring the Role of Design Task Choice on Learning, Design for Additive Manufacturing Use, and Creativity”, Journal of Mechanical Design, Vol. 142 No. 3, p. 031121, https://dx.doi.org/10.1115/1.4045127.

Reich, Y. (2022), “We cannot play 20 questions with creativity and innovation and win: the necessity of practice-based integrative research”, International Journal of Design Creativity and Innovation, Vol. 10 No. 2, pp. 69–74, https://dx.doi.org/10.1080/21650349.2022.2041889.

Reid, G.B. and Nygren, T.E. (1988), “The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload”, Advances in Psychology, Vol. 52, Elsevier, pp. 185–218, https://dx.doi.org/10.1016/S0166-4115(08)62387-0.

Rubio, S., Díaz, E., Martín, J. and Puente, J.M. (2004), “Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods”, Applied Psychology, Vol. 53 No. 1, pp. 61–86, https://dx.doi.org/10.1111/j.1464-0597.2004.00161.x.

Simpson, T.W., Boyer, J., Seepersad, C., Williams, C.B. and Witherell, P. (2016), “Special Issue: Designing for Additive Manufacturing”, Journal of Mechanical Design, Vol. 138 No. 12, https://dx.doi.org/10.1115/1.4034863.

Sorby, S.A. (1999), “Spatial Abilities And Their Relationship To Computer Aided Design Instruction”, presented at the ASEE Annual Conference and Exposition.

Tamhane, A.C. (1979), “A Comparison of Procedures for Multiple Comparisons of Means with Unequal Variances”, Journal of the American Statistical Association, Taylor & Francis, Vol. 74 No. 366a, pp. 471–480, https://dx.doi.org/10.1080/01621459.1979.10482541.

Tobias, S. and Carlson, J.E. (1969), “Brief Report: Bartlett’s Test of Sphericity and Chance Findings in Factor Analysis”, Multivariate Behavioral Research, Routledge, Vol. 4 No. 3, pp. 375–377, https://dx.doi.org/10.1207/s15327906mbr0403_8.

Tobii Pro AB. (2022), “Tobii Pro Lab User Manual (v1.194)”, Tobii Pro AB.

de Waard, D. and Lewis-Evans, B. (2014), “Self-report scales alone cannot capture mental workload”, Cognition, Technology & Work, Vol. 16 No. 3, pp. 303–305, https://dx.doi.org/10.1007/s10111-014-0277-z.

Wierwille, W.W. and Eggemeier, F.T. (1993), “Recommendations for Mental Workload Measurement in a Test and Evaluation Environment”, Human Factors: The Journal of the Human Factors and Ergonomics Society, Vol. 35 No. 2, pp. 263–281, https://dx.doi.org/10.1177/001872089303500205.