The Association between Physiological and Eye-tracking Metrics and Cognitive Load in Drivers: A Meta-analysis

Ange Wang[a,c], Chunxi Huang[b,d], Jiyao Wang[b], Dengbo He[a,b,c*]

[a] *Thrust of Intelligent Transportation, Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China*

[b] *Thrust of Robotics and Autonomous Systems, Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China*

[c] *Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China*

[d] *Robotics and Autonomous Systems, Division of Emerging Interdisciplinary Areas (EMIA) under Interdisciplinary Programs Office (IPO), The Hong Kong University of Science and Technology, Hong Kong SAR, China*

ABSTRACT

Driving performance can be impaired by a high cognitive load of drivers. Thus, it is important to estimate drivers' cognitive load. Although physiological and eye-tracking metrics have been widely used in many studies to assess cognitive load while driving, conflicts still exist regarding the association between physiological and eye-tracking metrics and different levels of cognitive load. Through a meta-analysis, our study aims to quantify the association between physiological, eye-tracking metrics and cognitive load induced by n-back tasks. A total of 18 articles met the inclusion criteria for the meta-analysis. The results indicate four types of metrics, including the sensitive-to-low ones that can only differentiate the low to medium level of cognitive load (i.e., the power spectrum of θ wave of electroencephalogram at Fp1 channel); high-resolution ones that can differentiate all levels of cognitive load (including pupil size, heart rate, and skin conductance); and low-resolution ones that can only differentiate low and high cognitive load (including the total power spectrum of electrocardiogram, eye blink rate, and respiration rate) and others (the power spectrum of θ wave of electroencephalogram at Fp2 channel). Furthermore, the association between metrics and cognitive load can be modulated by the n-back version, modality of n-back task, automation level, and percentage of male participants. In summary, this study contributes to the literature by quantifying associations between physiological and eye-tracking metrics and different cognitive load levels. Practically, we provide evidence for the selection of physiological and eye-tracking metrics for future driving cognitive load monitoring system design.

KEYWORDS

Physiological Metrics, Cognitive Load, Driving, N-back Task, Meta-analysis

1. INTRODUCTION

Human error is recognized as one of the dominating factors in road accidents (Singh, 2015). Though the human brain has long been recognized as a single-channel processor (C. D. Wickens, 1991), the driving task frequently involves multitasking. For example, during driving, in most cases, drivers need to control the vehicle (e.g., fine-tuning the gas pedal/brake pedal and the steering wheel to track the target speed and direction of the vehicle) and monitor the surrounding natural and traffic environment to identify potential hazards simultaneously. Different driving tasks may require different types of attention resources. Specifically, according to the multi-resource theory (Wickens, 1991), speed controlling task mainly requires visual-manual resources; while hazard perception tasks mainly require visual-cognitive resources. It has been commonly acknowledged that, compared to driving tasks that are visually and manually demanding, the cognitive demanding tasks, such as hazard perception and driving strategy selection are more safety-critical, and thus drivers' performance in these tasks is adopted as key metrics differentiating novice and experienced drivers (Jackson et al., 2009; Sagberg & Bjørnskau, 2006). In addition to driving-related tasks, in recent years, new technologies have been introduced into vehicles. For example, driving automation has been prevalent in newly sold vehicles, though it can reduce the overall workload of drivers, it may increase drivers' cognitive load because of the additional responsibility to monitor the automation (Stapel et al., 2019); while the introduction of infotainment functions in the smart cabin (e.g., video-streaming and internet browsing) and the prevalence of the bring-in smart devices (e.g., smartphones) may also increase the workload of drivers.

The high cognitive load in driving, either as a result of driving tasks or non-driving-related tasks, has been found to be closely related to driving safety in research environments (e.g., in driving simulators or instrumented vehicles). For example, a high cognitive load may lead to delayed responses to emergency events (Harbluk et al., 2007), reduced visual search scope, leading to a visual tunnel effect (Recarte & Nunes, 2000)), decreased ability to anticipate hazards (Muhrer & Vollrath, 2011), and increased reaction times (Du et al., 2020) and impaired performance (Melnicuk et al., 2021) in takeover events during assisted driving. Thus, estimating drivers' high cognitive load can be a potential approach to improve driving safety, both in vehicles with and without driving automation.

As an intrinsic state, cognitive load can hardly be measured objectively and directly (e.g., the questionnaire is direct but subjective, while eye-tracking measures are objective but indirect). Moreover, unlike distracted driving and fatigue, the state of high cognitive load is not easily discernible from the normal driving state, as it can be an integral part of the driving task. In the domain of driving, the cognitive load can be evaluated using four different types of measures, i.e., subjective measures, such as the NASA-Task Load Index (NASA-TLX) scale (Hart & Staveland, 1988); physiological measures, such as the electrocardiogram (ECG), electroencephalogram (EEG), respiration, and electrodermal activity (EDA); eye-tracking measures, such as pupil size;

and task performance measures, including driving task measures and non-driving-related task measures. All these measures have their pros and cons. For example, the subjective questionnaire methods cannot estimate cognitive load in real-time. The non-driving-related tasks can disrupt drivers' natural driving behavior and increase their cognitive load. The driving task measures are highly susceptible to traffic conditions and become invalid during automated driving, given that drivers do not need to control the vehicle for extended periods. Thus, the physiological measures and eye-tracking measures are most promising for real-time cognitive load detection. Further, with the advancement of new technologies, non-intrusive measures of physiological metrics and highly accurate eye-tracking measures have become possible (Ayres, 2020).

However, although associations between some physiological and eye-tracking metrics and the variations in cognitive load have been observed in some studies, no consensus has been reached for some other physiological and eye-tracking metrics, which poses challenges in selecting appropriate metrics for developing cognitive load detection algorithms in drivers. For example, the correlation between the respiration rate (RR) and cognitive load has been found to be negative in some studies (He et al., 2019), but positive in some other studies (Hajek et al., 2013). For other metrics, substantial differences in the strength of the association have been identified. For example, in Rahman et al (2020), the Low Frequency (LF) power (0.04–0.15 Hz) of heart rate variability (HRV) exhibited a strong positive correlation ($r > 0.8$) with cognitive load. However, only a weak correlation ($r < 0.1$) has been found in Tjolleng et al (2017). Lastly, not all metrics are responsive in differentiating different levels of cognitive load. For example, the heart rate (HR) was able to differentiate between median to high levels of cognitive load but showed no difference between low to median levels of cognitive load (Ferreira et al., 2014). However, the feasible range of different cognitive load measures has not been systematically analyzed, which hinders the development of different algorithms targeting different levels of cognitive load, using minimum types of measures.

Thus, it is necessary to quantify the relationships between the physiological and eye-tracking metrics and the driver's cognitive load levels. Given that not all metrics are responsive to the whole range of cognitive load, the meta-analysis needs to be conducted for different ranges of cognitive load levels and a metric or task that can consistently impose different levels of cognitive load to drivers has to be selected. Traditionally, subjective responses such as NASA-TLX were regarded as the ground truth of cognitive load levels in previous studies (Chen et al., 2022; He et al., 2019; Hart & Staveland, 1988). Although NASA-TLX allows within-subject comparisons, individual differences in self-reported scores may exist and we can hardly compare the NASA-TLX scores across participants and experiments (Muth et al., 2012). Thus, in this study, we adopted a more standard task to label the levels of cognitive – the n-back task.

The n-back task is mainly a working memory task and has been proven as an effective manipulation of cognitive load in vehicles (Mehler et al., 2012a; von

Janczewski et al., 2021). In the n-back task, a series of stimuli (such as numbers or letters) are presented. Between each stimulus, there is a sustained pause that allows participants to repeat the stimuli presented *n* positions before (see an example in Figure 1). The levels in the n-back task (i.e., *n*) reflect the difficulty and complexity of the task, the larger the *n*, the more difficult the task is. The utilization of 3-back is rare in research as 2-back is already sufficiently demanding. A substantial body of behavioral and neuroimaging research has confirmed the sensitivity of different levels of N-back tasks to cognitive load (Broadbent et al., 2023; Rieck et al., 2022; Solhjoo et al., 2019). Further, other commonly used cognitive tasks usually have no clear standard definition of cognitive levels. For example, the difficulty levels of hybrid tasks (obstacle avoidance and recall) (Yang et al., 2023) and mathematical tasks (von Janczewski et al., 2021) are difficult to quantify (e.g., the difficulty level of “3455 – 15” and “3455 – 7” are not clearly quantified, though the former has one more digit than the latter). Given that the difficulty of 0-, 1- and 2-back tasks are comparable to common in-vehicle tasks (Mehler & Reimer, 2019) in our study, we treat the cognitive level triggered by the driving task only and 0-back task as low, the cognitive load triggered by 1-back task as median, and the cognitive load triggered by 2-back task as high.

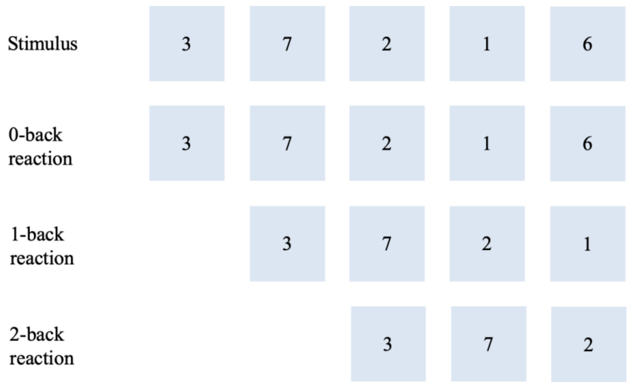

Fig. 1. Demonstration of a typical n-back task procedure for n = 0, n = 1, and n = 2.

Therefore, in this study, based on a meta-analyses approach, we systematically analyzed the changes in drivers' physiological and eye-tracking metrics in response to the variation in cognitive load during driving, as defined by the levels of the n-back tasks. All metrics explored in this study were found to be associated with the cognitive load at least in some of the previous studies. To the best of our knowledge, though previous meta-analysis validated the effectiveness of n-back task in imposing high cognitive load in drivers (von Janczewski et al., 2021), no research has quantified the relationship between drivers' cognitive load and their physiological and eye-tracking measures. The summary of the abbreviations, descriptions, and units of the physiological and eye-tracking metrics mentioned in the text is shown in Table 1.

In addition, given that the cognitive load is a multi-dimensional concept and the settings in different studies can affect the responses of physiological and eye-tracking measures (Nilsson et al., 2022). we adopted meta-regression to account for 1) the influence of measurable individual differences so that future driver monitoring systems may take adaptive strategies; 2) and artificial experiment settings so that we can better isolate the cognitive effects. Specifically, the demographic variables, including age (Wickens et al., 2011; Zhang et al., 2017) and gender (Sârbescu et al., 2014; Zhang et al., 2017) that can affect driving behavior, and the experiment-related settings including simulator fidelity, experimental environment, the modality of stimulus presentation and response, and the time interval between stimuli (Janczewski et al., 2021) that could modulate the drivers' cognitive load were considered. Besides, different versions of n-back tasks were used in previous driving studies. For example, the participants were required to only memorize and repeat the item that is n position before (**repeating version**, see Figure 1) in Gable et al (2015) and Mehler et al (2012b); while in Nilsson et al (2022) and Rahman et al (2020), participants needed to indicate whether the judge if the two items that are n positions apart are the same or not (**matching version**), which required additional cognitive resource to compare the two items. Another version of the n-back task was used by He et al (2019), which required participants to count how many times a pattern appeared (**counting version**) in addition to the matching version. Different versions require different cognitive components and hence the n-back version is also considered to account for the multi-dimensionality of the cognitive task.

In summary, the contribution of this study is 2-fold. First, the meta-analysis approach was employed to investigate the association between physiological, and eye-tracking metrics and cognitive load levels. Second, the moderating effects of participants' age and gender, driving automation level, the fidelity of simulators, n-back version, modality of n-back stimulus and response, and the time interval between stimuli on these associations were explored.

Table 1. Summary of the abbreviations, descriptions, and units of the physiological metrics.

Measure	Metric	Abbreviation	Description	Unit
Eye	Pupil size	PS	The diameter of the opening in the center of the iris	mm
	Fixation duration	FD	The length of time that gaze remains focused on a specific object or region of interest	ms
	Eye blink rate	EBR	The number of blinks per unit of time	blinks/min
	Eyeblink duration	EBD	The time taken for each blink from closure to reopening of the eyes	ms
Electroencephalography (EEG)	Theta wave-Fp1 channel	θ-Fp1	The θ waves (4-8 Hz) located in the Fp1 channel.	μV
	Theta wave-Fp2 channel	θ-Fp2	The θ waves (4-8 Hz) located in the Fp2 channel.	μV
Electrocardiogram (ECG)	Heart rate	HR	The number of heartbeats occurring per minute	Beats/minute
	Standard deviation of normal-to-normal intervals	SDNN	The variability of the time intervals between consecutive normal heartbeats	ms
	Root mean square of successive differences	RMSSD	The magnitude of the differences between consecutive R-R intervals (the time between successive heartbeats)	ms
	Low frequency	LF	The spectral power in the low-frequency range (usually 0.04 to 0.15 Hz) of the heart's electrical activity	ms^2
	High frequency	HF	The spectral power in the high-frequency range (usually 0.15 to 0.4 Hz) of the heart's electrical activity	ms^2
	Low frequency/ High frequency	LF/HF	Ratio of LF to HF	%
	Total power	TP	The overall power spectrum	ms^2
	pNN50	pNN50	The percentage of successive RR intervals (the time between R-peaks on an ECG) that differ by more than 50 milliseconds	%
	Very low frequency.	VLF	The frequency range of electrical signals in the ECG waveform that are below 0.04 Hz	Hz
	Low frequency power	LFun	The power or intensity of low-frequency electrical activity	ms^2

	High frequency power	HFun	The power or intensity of high-frequency electrical activity	ms²
	Inter-beat interval	IBI	The time duration between successive heartbeats	ms
Skin	Electrodermal activity	SC-EDA	The general term that encompasses the electrical activity of the skin	μS
Respiration	Respiration rate	RR	The number of breaths taken per minute	Respirations / minute

2. METHOD

2.1 Literature search and study selection

We adopted the approach based on the PRISMA statement (Moher, 2009), a comprehensive guideline for reporting items in meta-analysis. An extensive literature search was conducted, covering articles published up until Feb 2024. The study selection process is summarized in Figure 2. In our search, the title, abstract or keywords must include ("driver" OR "driving" OR "automobile" OR “automated” OR "vehicle" OR "car") and (“cognitive load” OR “workload” OR “working memory” OR “mental workload”); and the full text must have "N-back" and (“physiological” OR “eye” OR “electroencephalogram” OR “electrocardiogram” OR “respiration” OR “electrodermal activity” OR “galvanic skin reaction” OR “skin conductance” OR “pulse rate variability” OR “temperature”).

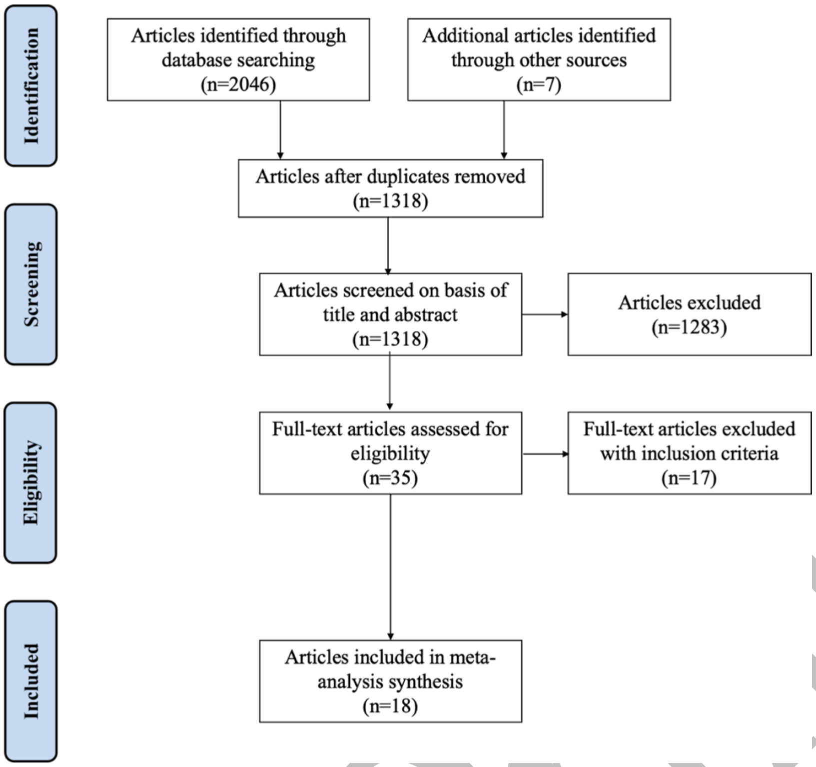

Fig. 2. Literature review process based on the PRISMA method.

2.2 Inclusion and exclusion criteria

The inclusion criteria for studies were as follows: (1) The n-back task should be a task secondary to the driving task of four-wheel cars and only empirical studies conducted on real roads or in simulated driving environments were included. (2) The study should have at least two different n-back levels or one n-back level and a baseline without a secondary task. (3) The statistics of the physiological or eye-tracking measures must be reported or could be obtained by contacting the authors. (4) The physiological and eye-tracking measures associated with the n-back level should be independent of other tasks in the vehicle so that the cognitive load was only induced by the n-back task. (5) Due to language barriers, only publications in Chinese and English were included. Publications that did not meet any of the above criteria were excluded from the analysis. Initial determinations were made based on the abstracts, followed by a thorough examination of the full texts based on the inclusion criteria. Ultimately, 18 studies were kept for further analysis.

2.3 Data extraction

The following information was extracted from each publication: (1) Meta-information of the study (i.e., title, author, and publication year); (2) Descriptions and measures of the cognitive workload; (3) sample characteristics (i.e., sample size, mean age, and characteristics of participants); (4) experimental conditions (i.e., automation level, experimental environment, simulator fidelity); (5) characteristics of n-back tasks (i.e., modality of the stimulus and response, and time interval between stimuli) and (6) associations between the physiological and eye-tracking metrics and n-back levels or the raw values of the metrics under each experimental condition.

These data were extracted and coded independently by two doctoral students (the first two authors). To ensure a consistent understanding of the coding scheme between the two coders, we conducted a preliminary "coding trial" phase. During this phase, the two coders independently coded five articles and discussed any discrepancies in coding to reach a consensus on the coding scheme. Necessary modifications and refinements were made to the coding manual based on the issues encountered during this phase.

2.4 Data processing

In the meta-analyses, the Pearson correlation coefficient (*r*) was used as the effect size for each study. We followed a systematic eight-step process to analyze the data for our meta-analyses. An overview of these steps is provided in Figure 3. In the meta-analyses, we focused on conducting meta-analyses of the physiological and eye-tracking metrics. Specifically, the physiological and eye-tracking metric from each level of cognitive load (as imposed by the n-back task or baseline task) was compared to all other levels of cognitive load. In total, six (C_3^2=3 for baseline and 2 levels of n-back tasks) pairwise comparisons for all metrics were explored in this study. If the correlation coefficient was not provided in the study, we did the calculation according to Lipsey & Wilson (2001) based on the mean value, standard deviation (SD), and sample size. It is worth noting that the meta-analyses were only conducted for metrics that were investigated in at least two studies, which is the minimum requirement for a meta-analysis (McCarthy et al., 2017; Zheng, 2013).

Then, for each metric in each pairwise comparison, we transferred r (ranging between -1 and 1) to Fisher's Z value following the equation in Lipsey & Wilson (2001):

$$Z_r = 0.5 * \ln((1 + r) / (1 - r)) \quad (1)$$

where Z_r represents Fisher's Z value, and r denotes the correlation coefficient between two variables. This transformation can alleviate the constraints of the correlation coefficient, as Z_r ranges from negative infinite to positive infinite, which enables the weighted combination of effect sizes from multiple studies in meta-analysis.

Next, the meta-analyses were conducted using RevMan 5.4 (Schmidt et al., 2019). The random-effects model was used to calculate the weighted average correlation, in which the calculation of weight factors was based on the inverse of the variance (Lipsey

& Wilson, 2001). The forest plots were provided to visualize the results (see Appendix 1), in which, the pooled Z_r derived from the amalgamation of all incorporated studies was visually represented as a rhombus positioned at the lower section of the graph, wherein the breadth of the rhombus denoted the 95% confidence interval (95%CI). The significance of the estimated overall effect size was evaluated using the 95%CI (Lipsey & Wilson, 2001) and the p-value, with .05 as the significance threshold.

In addition, for each meta-analysis, the heterogeneity of the research was evaluated using the I^2 statistic, tau squared (τ^2), and Q statistic (Lipsey & Wilson, 2001). The I^2 statistic quantifies the proportion of the observed variation in correlation that can be accounted for by actual variations between studies. A value of 25%, 50%, or 75% corresponds to low, moderate, or high levels of variance, respectively (Higgins & Deeks, 2003). The τ^2 represents the overall extent of heterogeneity, with a smaller τ^2 indicating a lower level of heterogeneity. The Q statistic reflects the degree of heterogeneity resulting from actual differences between studies, with a significant Q statistic implying the existence of heterogeneity among the studies. Out of the above-mentioned metrics, the Q-test is the most used for testing heterogeneity. However, its testing power is limited when the number of studies is small. In contrast, the I^2 statistic can mitigate the impact of the sample size on the testing power. In our study, we adopted the I^2 of 50% as the threshold for the existence of heterogeneity (Zheng, 2013), but still reported τ^2 and Q statistics for readers' reference.

Finally, to investigate the origins of research heterogeneity, meta-regression was conducted for potential moderators, including experimental conditions (i.e., automation level), characteristics of n-back tasks (i.e., modality of the stimuli and responses, the time interval between stimuli), and demographic variables (i.e., mean age and percentage of male). The automation level here refers to driving automation defined by the Society of Automotive Engineers (SAE) (Committee, 2014). It is worth noting that we refined the simulator-type classification scheme proposed by Spyridakos et al., (2020) as follows: "Occlusion" and "Desktop" were classified as the category of low-fidelity driving simulators, "Cabin with narrow field projection" and "Cabin with widefield projection" were classified as medium fidelity driving simulators; and "Hexapod," "Hexapod and lateral motion," and "Hexapod and longitudinal motion" were classified as high-fidelity driving simulators. Furthermore, since the median of the median age among the 18 studies included in the analysis is 27.2, we thus adopt the threshold of 27 years for age stratification.

Then, intergroup homogeneity was performed and heterogeneity coefficients were computed to assess the intergroup effects (Viechtbauer, 2010). For significant moderators, sub-group meta-analyses were conducted. Subsequently, we proceeded to assess the homogeneity of effect sizes within a specific group (Q_W) and the heterogeneity across different sub-groups (Q_B) (Lipsey & Wilson, 2001). The moderator analyses were conducted using Stata 17.

It should be noted that to facilitate interpretation, when reporting the results, the Z_r was transformed back to r. According to Cohen (1988), $|r| \leq 0.3$ denotes a small correlation, $0.3 \leq |r| \leq 0.5$ denotes a medium correlation, and $|r| \geq 0.5$ denotes a large correlation.

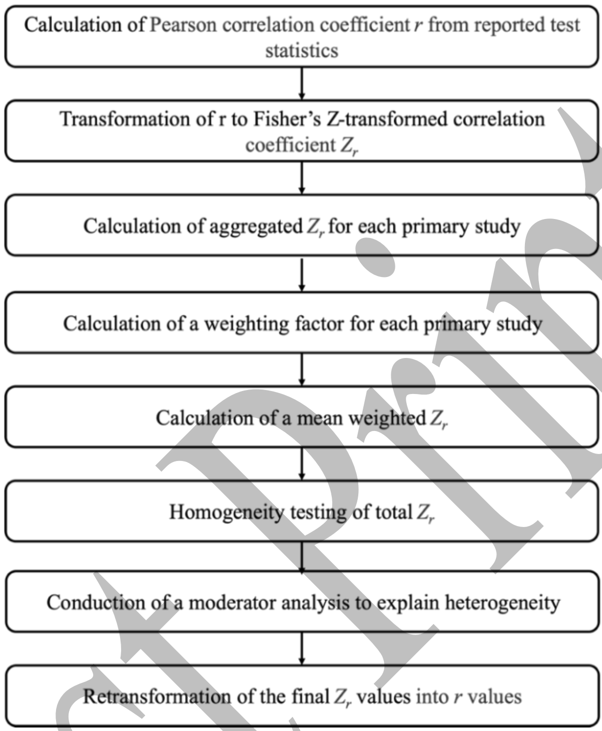

Fig. 3. Overview of the Processing Workflow.

3. RESULT

The process of literature retrieval and study selection is illustrated in Figure 3. Initially, a total of 2,046 records were retrieved through database search, and an additional 7 records were included after examining the reference list of the identified literature. After removing duplicates, 1,318 records remained, which were then screened based on their titles and abstracts. Then, full-text evaluation was conducted for 35 records, and 18 records that met our criteria were kept for meta-analyses.

1 **Table 2**. A summary of the identified literature

Study	Participant Number	Metrics	Automation Level	Experimental Environment	N-back Levels	N-back Version	N-back Modalities (stimuli-responses)	Inter-stimulus Interval (s)	Male Ratio (%)	Mean Age (SD/Range)
(Deng et al., 2024)	20	HR, SC-EDA	L3	Medium fidelity simulator	N, 1B, 2B	Matching	Visual-verbal	—	90	25.3 (3.8/23-39)
(Nilsson et al., 2022)	70	HR, RMSSD, RR, SC-EDA, PS, EBR, EBD	L0	Medium fidelity simulator	N, 1B, 2B	Matching	Auditory-manual	1	100	43 (4/35-51)
(Mehler et al., 2012b)	108	HR, SC-EDA	L0	On-road	N, 0B, 1B, 2B	Repeating	Auditory-verbal	—	50	24.6 (2.7/-), 44.5 (3.0/-), 63.3 (3.1/-)
(Chen et al., 2022)	36	PS	L0, L1, L2	Low fidelity simulator	N, 2B	Repeating	Visual-verbal; Auditory-verbal; Auditory-spatial	0.2	50	26 (2.8/-)
(Zhang et al., 2022)	18	θ-Fp1, θ-Fp2, SC-EDA, IBI, HR, SDNN, RMSSD, LF, HF, LF/HF	L0	Low fidelity simulator	N, 1B, 2B	Matching	Auditory-manual	—	100	23.2 (1.6/-)
(Meteier et al., 2022)	80	HR	L3	Low fidelity simulator	N, 1B, 3B	Matching	Auditory-manual; Visual-manual	0.5	32.5	23.9 (8.2/19-66)
(Yang et al., 2021)	75	HR, SDNN, TP	L0	Medium fidelity simulator	N, 2B	Matching	Auditory-manual	1	67	31 (11.6/-)
(Gable et al., 2015)	8	HR, PS	L0	Low fidelity simulator	N, 0B, 1B, 2B	Repeating	Auditory-verbal	2.25	62.5	21.1 (19-23/-)
(Du et al., 2020)	102	HR	L3	High fidelity simulator	1B, 2B	Matching	Visual-manual	2.5	—	22.9 (3.8/18-38)
(Rahman et al., 2020)	33	SDNN, RMSSD, pNN50, VLF, LF,	L0	High fidelity simulator	N, 1B, 2B	Matching	Auditory-manual	—	100	42.5 (-/35-50)

		HF, TP, LF/HF, LFun, HFun								
(Cegovnik et al., 2018)	22	PS	L0	Low fidelity simulator	N, 0B, 1B, 2B, 3B	Repeating	Auditory-verbal	5.5	81.8	32.9 (-/22-61)
(Tjolleng et al., 2017)	15	IBI, LF	L0	Medium fidelity simulator	0B, 1B, 2B	Repeating	Auditory-verbal	—	100	27.7 (3.0/-)
(Niezgoda et al., 2015)	46	PS	L0	High fidelity simulator	N, 0B, 1B, 2B	Repeating	Auditory-verbal	2.25	63	36.6 (9.7/21-59)
(Reimer et al., 2009)	26	HR, EDA	L0	On-road	N, 0B, 1B, 2B	Repeating	Auditory-verbal	—	—	23.9 (1.6/-)
(Hajek et al., 2013)	47	HR, RR, SC-EDA, RMSSD	L2	Low fidelity simulator	N, 2B	Repeating	Auditory-verbal	2.25	72.3	28.5 (8.7/19-55)
(Zheng et al., 2021)	20	SDNN, RMSSD, pNN50, LF, HF, TP, LF/HF, VLF, LFun, HFun, SC-EDA	L0	Low fidelity simulator	N, 0B, 1B, 2B	Repeating	Auditory-verbal	2.5	75	26.7 (3.8/-)
(Mehler et al., 2009)	121	HR, SC-EDA, RR	L0	Medium fidelity simulator	N, 0B, 1B, 2B	Repeating	Auditory-verbal	2.25	48.8	24.5 (2.8/20-29)
(He et al., 2019)	34	PS, HR, RR, SC-EDA, θ-Fp1, θ-Fp2	L0	Low fidelity simulator	N, 1B, 2B	Counting	Auditory-verbal	2.5	48.6	26.4 (4.3/20-35)

Note: "—" means that the information has not been mentioned in the corresponding paper. In the table, N, 0B, 1B, 2B, and 3B standard for baseline without n-back task, 1-back task, 2-back task, and 3-back task, respectively. L0, L1, L2, and L3 donate SAE Level 0, Level 1, Level 2, and Level 3 automation, respectively.

3.1 Descriptive statistics

Table 2 provides descriptive information of all studies included in the meta-analyses. Overall, a total of 881 participants were involved in the experiments and experienced various levels of n-back tasks. Table 3 provides a summary of correlation coefficients between the metrics and cognitive load levels in all studies included for meta-analyses. It was found that cognitive levels induced by 1-back and 2-back tasks were most intensively investigated. At the same time, among all psychological and eye-tracking measures, the relationship between heart measures and cognitive load levels attracted the most attention in previous research, with HR attracting the most attention among heart-related metrics. Moreover, the total sample sizes of studies for a single meta-analysis varied widely, ranging from 38 to 711 participants. Finally, it should be noted that substantial differences in correlation coefficients of the metrics have been observed between different studies, confirming the need for further meta-analyses.

At the same time, given the small number of studies that could be identified, and considering the task difficulties, the 0-back (which only requires participants to simply repeat what they hear immediately) and baseline without n-back were aggregated as low task load (L); the 1-back was labeled as medium task load (M); and the 2-back was categorized as high task load (H).

Table 3. Correlation coefficients between the physiological and eye-tracking metrics and cognitive load levels in all studies.

Metrics			*n*	Sample size			*r*			
				Min	Max	Total	Mean	SD	Min	Max
L vs. M	Eye	PS	6	8	70	212	0.39	0.31	-0.12	0.77
		EBR	3	46	70	162	0.11	0.03	0.08	0.15
	Brain	θ-Fp1	2	18	34	52	0.42	0.04	0.40	0.45
		θ-Fp2	2	18	34	52	0.39	0.05	0.36	0.43
	Heart	HR	18	8	121	711	0.22	0.19	-0.2	0.76
		SDNN	3	18	33	51	0.35	0.65	-0.34	0.94
		RMSSD	3	18	33	51	0.36	0.65	-0.34	0.94
		LF	5	18	33	106	0.20	0.44	-0.13	0.97
		HF	3	18	33	71	0.34	0.45	-0.22	0.88
		LF/HF	3	18	33	71	0.49	0.46	0.05	0.97
		TP	2	20	33	53	0.35	0.87	-0.26	0.97
		pNN50	2	20	33	53	0.42	0.79	-0.14	0.98
		VLF	2	20	33	53	0.47	0.51	-0.05	0.98
		LFun	2	20	33	53	0.50	0.70	0.00	0.99
		HFun	2	20	33	53	0.35	0.90	-0.29	0.99
	Skin	EDA	15	18	121	571	0.15	0.26	-0.52	0.8
	Respiration	RR	3	34	121	225	0.13	0.34	-0.34	0.44
L vs. H	Eye	FD	2	36	46	82	0.02	0.22	-0.14	0.18
		PS	7	8	46	248	0.37	0.34	-0.2	0.82
		EBR	4	36	70	198	0.12	0.12	0.12	0.22
		EBD	2	36	70	106	-0.09	0.11	-0.11	-0.08
	Brain	θ-Fp1	2	18	34	42	0.33	0.02	0.32	0.35
		θ-Fp2	2	18	34	42	0.28	0.33	0.04	0.51
	Heart	HR	17	8	121	653	0.47	0.30	0.00	0.69
		SDNN	3	18	75	113	-0.23	0.58	-0.70	0.43
		RMSSD	3	18	47	85	0.30	0.61	-0.37	0.84
		LF	4	18	33	73	0.03	0.19	-0.22	0.30
		HF	4	15	20	73	0.04	0.18	-0.18	0.30
		LF/HF	2	18	20	38	0.27	0.15	0.16	0.38
		TP	2	20	75	95	-0.34	0.19	-0.21	-0.48

	Skin	EDA	14	18	121	628	0.11	0.33	-0.99	0.45
	Respiration	RR	4	34	121	272	0.31	0.27	-0.13	0.59
M vs. H	Eye	PS	4	8	46	158	0.23	0.10	0.13	0.34
		EBR	2	46	70	116	0.08	0.02	0.06	0.1
	Brain	θ-Fp1	2	18	34	52	0.02	0.02	0.01	0.04
		θ-Fp2	2	18	34	52	-0.25	0.02	-0.27	-0.24
		HR	12	8	121	522	0.17	0.16	0.02	0.56
		SDNN	3	18	33	71	0.30	0.52	-0.09	0.84
		RMSSD	4	18	70	141	0.21	0.40	-0.19	0.84
		LF	4	15	33	86	0.34	0.38	0.04	0.98
		HF	3	18	33	71	0.33	0.36	0.02	0.84
		LF/HF	3	18	33	71	0.46	0.44	0.12	0.96
		TP	2	20	33	53	0.50	0.63	0.06	0.95
		pNN50	2	20	33	53	0.40	0.74	-0.12	0.92
		VLF	2	20	33	53	0.51	0.45	0.06	0.96
		LFun	2	20	33	53	0.52	0.65	0.05	0.98
		HFun	2	20	33	53	0.41	0.82	-0.18	0.99
	Skin	SC-EDA	11	18	121	437	0.23	0.27	0.00	0.80
	Respiration	RR	3	34	121	225	0.08	0.08	0.02	0.19

Note："vs." denotes the act of conducting pairwise comparisons, *n* denotes the number of studies included.

3.2 Results for meta-analyses

The results of the meta-analyses are presented in Table 4. A majority of the analyses in Table 4 have a significant unexplained variance (I^2>50%), indicating the necessity for moderators' analyses. The forest plots of the meta-analysis in Table 4 are provided in Appendix 1.

Table 4. Meta-analyses' results of the association between the physiological metrics and cognitive load levels.

Metrics			*k*	Random Effects Model				Heterogeneity Test		
				n	Pooled Z_r (95 % CI)	Pooled *r* (95 % CI)	*p*	τ^2	Chi^2 (*df*)	I^2 (%)
L vs. M	Eye	PS	212	6	0.48 (0.13,0.82)	0.45 (0.13, 0.68)	**.006**	0.15	38.82 (5)	**87**
		EBR	162	3	0.11 (-0,0.22)	0.11 (-0,0.22)	.06	0	0.23 (2)	0
	Brain	θ-Fp1	52	2	0.43 (0.21,0.66)	0.41 (0.21, 0.58)	**<.001**	0.00	0.06 (1)	0
		θ-Fp2	52	2	0.44 (0.22,0.67)	0.41 (0.22, 0.58)	**<.001**	0.00	0.1 (1)	0
	Heart	HR	711	18	0.22 (0.13,0.32)	0.22 (0.13, 0.32)	**<.001**	0.02	38.92 (17)	**56**
		SDNN	51	3	0.88 (-0.06,1.82)	0.71 (-0.06, 0.95)	.07	0.63	26.61 (2)	**92**
		RMSSD	121	3	0.61 (-0.34,1.56)	0.61 (-0.33, 0.92)	.2	0.9	91.33 (3)	**97**
		LF	106	5	0.43 (-0.48,1.33)	0.41 (-0.45, 0.87)	.4	1.02	109.35 (4)	**96**
		HF	71	3	0.52 (-0.47,1.51)	0.48 (-0.44, 0.91)	.3	0.71	29.62 (2)	**93**
		LF/HF	71	3	0.89 (-0.43,2.21)	0.71 (-0.41, 0.98)	.2	1.31	53.86 (2)	**96**
		TP	53	2	0.92 (-1.40,3.23)	0.73 (-0.89, 1.00)	.4	2.74	60.76 (1)	98
		pNN50	53	2	1.09 (-1.31,3.48)	0.80 (-0.86, 1.00)	.4	2.93	64.13 (1)	98
		VLF	53	2	1.13 (-1.17,3.43)	0.81(-0.82,1)	.3	2.71	59.51(1)	98
		LFun	53	2	1.33 (-1.27,3.93)	0.87 (-0.85, 1.00)	.3	3.46	75.66 (1)	99
		HFun	53	2	1.18 (-1.7,4.07)	0.83 (-0.94, 1.00)	.4	4.27	92.11 (1)	99
	Skin	SC-EDA	551	14	0.20 (0.08,0.31)	0.20 (0.08,0.30)	**<.001**	0.03	37.92(13)	66
	Respiration	RR	225	3	0.15 (-0.28,0.58)	0.15 (-0.27,0.52)	.5	0.13	30.29 (2)	93
L vs. H	Eye	FD	82	2	0.03 (-0.28,0.34)	0.03 (-0.27, 0.33)	.9	0.02	1.93 (1)	48
		PS	248	7	0.47 (0.18,0.76)	0.44 (0.18, 0.64)	**.002**	0.12	49.12 (6)	**88**
		EBR	198	4	0.16 (0.06,0.27)	0.16 (0.06,0.26)	**.002**	0	2.44 (3)	**0**
		EBD	106	2	-0.09 (-0.23,0.06)	-0.09 (-0.23,0.06)	.3	0	0.02 (1)	**0**
	Brain	θ-Fp1	52	2	-0.05 (-0.72,0.62)	-0.05 (-0.62, 0.55)	.9	0.2	5.73 (1)	83
		θ-Fp2	52	2	-0.35 (-1.06,0.36)	-0.34 (-0.79, 0.35)	.3	0.22	6.12 (1)	84
	Heart	HR	653	17	0.4 (0.25, 0.54)	0.38 (0.24, 0.49)	**<.001**	0.07	83.37 (16)	**81**
		SDNN	113	3	-0.3 (-1.06,0.46)	-0.29 (-0.79, 0.43)	.4	0.41	22.09 (2)	**91**
		RMSSD	155	4	0.24 (-0.79,1.27)	0.24 (-0.66, 0.85)	.7	1.06	248.43 (3)	**99**
		LF	73	4	0.02 (-0.17,0.21)	0.02 (-0.17,0.21)	.8	0	2.54 (3)	**0**
		HF	73	4	0.03 (-0.17,0.22)	0.03 (-0.17,0.22)	.8	0	2.2 (3)	**0**

		LF/HF	38	2	0.28 (-0.07,0.63)	0.27 (-0.07, 0.56)	.1	0	0.42 (1)	0
		TP	95	2	-0.47 (-0.68, -0.26)	-0.44 (-0.59, -0.25)	**<.001**	0	0.86 (1)	0
	Skin	SC-EDA	608	13	0.20 (0.12,0.28)	0.20 (0.12,0.28)	**<.001**	0	15.27 (12)	21
	Respiration	RR	272	4	0.34 (0.04,0.63)	0.34 (0.04, 0.56)	**.03**	0.08	23.3 (3)	**87**
M vs. H	Eye	PS	158	4	0.29 (0.17,0.41)	0.28 (0.17, 0.39)	**<.001**	0	2.15 (3)	0
		EBR	116	2	0.08 (-0.05,0.22)	0.08 (-0.05, 0.22)	.2	0	0.08 (1)	0
	Brain	θ-Fp1	52	2	0.03 (-0.16,0.22)	0.03 (-0.16, 0.22)	.8	0	0.02 (1)	0
		θ-Fp2	52	2	-0.25 (-0.45, -0.06)	-0.24 (-0.42, -0.06)	**.009**	0	0.04 (1)	0
		HR	522	12	0.14 (0.05,0.23)	0.14 (0.05, 0.23)	**.002**	0.01	18.74 (11)	41
		SDNN	71	3	0.47 (-0.42,1.36)	0.44 (-0.40, 0.88)	.3	0.59	43.72 (2)	**95**
		RMSSD	141	4	0.29 (-0.21,0.78)	0.29 (-0.21, 0.65)	.3	0.24	52.57 (3)	**94**
		LF	86	4	0.67 (-0.35,1.69)	0.59 (-0.34, 0.93)	.2	1.06	110.6 (3)	**97**
		HF	71	3	0.45 (-0.26,1.17)	0.42 (-0.25, 0.82)	.2	0.37	28.15 (2)	**93**
		LF/HF	71	3	0.79 (-0.3,1.88)	0.66 (-0.29, 0.95)	.2	0.91	65.5 (2)	**97**
		TP	53	2	0.94 (-0.79,2.68)	0.74 (-0.66, 0.99)	.3	1.54	53.36 (1)	98
		pNN50	53	2	0.73 (-0.94,2.41)	0.62 (-0.74, 0.98)	.4	1.43	49.7 (1)	98
		VLF	53	2	1 (-0.85,2.85)	0.76 (-0.69, 0.99)	.3	1.76	61.34 (1)	**98**
		LFun	53	2	1.35 (-1.2,3.9)	0.87 (-0.83, 1.00)	.3	3.35	117.52 (1)	99
		HFun	53	2	1.23 (-1.54,4.01)	0.84 (-0.91, 1.00)	.4	3.98	137.67 (1)	99
	Skin	SC-EDA	417	10	0.18 (0.04,0.33)	0.18 (0.04,0.33)	**.01**	0.04	35.95 (9)	**75**
	Respiration	RR	225	3	0.05 (-0.04,0.14)	0.05 (-0.04, 0.14)	.3	0	1.53 (2)	0

Notes：L denotes the baseline condition and 0-back task, 1B denotes the 1-back task, and 2B represents the 2-back task, whereas "vs." denotes the act of conducting pairwise comparisons, *k* denotes the cumulative sample size, *n* denotes the number of studies included. The bolded pooled *r* indicates significant (p<.05) metrics. The bolded I^2 indicates the existence of heterogeneity of the metrics.

3.3 Moderator analysis

Given that the inclusion of a sufficient number of studies is required for conducting meta-regression, we specifically focused on meta-analyses with an I^2 greater than 50% or *p*-value in heterogeneity tests smaller than .05 and a minimum of three included studies. The results of the meta-regression model are summarized in Table 5. In addition, inter-group homogeneity tests were conducted on significant moderating factors ($p<.05$). Table 6 displays the weighted average effect size Z_r and r (as well as their 95%CI) for each subgroup, as well as the Q_W value that captures the overall heterogeneity within all the sub-groups of one moderator. Additionally, Q_B values are listed for each moderating variable, indicating the presence of heterogeneity among subgroups for each moderating factor (Lipsey & Wilson, 2001). The forest plots regarding the aggregated Z_r for each subgroup analysis are provided in Appendix 2.

Table 5. Results of meta-regression models.

	Moderators	L vs. M			L vs. H			M vs. H		
		n	*Coefficient*	*p*	*n*	*Coefficient*	*p*	*n*	*Coefficient*	*p*
PS	Simulator fidelity	6	0.20	.3	7	0.19	.4	-	-	-
	Modality of stimuli and responses	6	0.04	.95	7	-0.03	.9	-	-	-
	n-back version	**6**	**-0.46**	**.02**	7	-0.14	.6			
	Inter-stimulus interval	6	0.13	.8	7	-0.03	.9			
	Percentage of males	6	0.001	.09	7	0.01	.3	-	-	-
	Mean age	6	0.01	.7	7	0.04	.07	-	-	-
HR	Automation level	18	0.09	.09	**17**	**0.33**	**<.0001**	-	-	-
	Experimental environment	18	0.06	.6	17	0.32	.06	-	-	-
	Simulator fidelity	10	0.17	.5	9	-0.50	.14	-	-	-
	n-back version	18	0.12	.2	17	0.20	.2			
	Modality of stimuli and responses	18	0.08	.09	**17**	**0.18**	**.04**	-	-	-
	Inter-stimulus interval	8	0.10	.08	7	0.16	.4	-	-	-
	Percentage of males	16	0.002	.4	**15**	**0.01**	**.02**	-	-	-
	Mean age	18	-0.003	.4	17	-0.009	.2	-	-	-
EDA	Automation level	**14**	**0.22**	**<.0001**	-	-	-	**10**	**0.21**	**.005**
	Simulator fidelity	6	0.15	.7	-	-	-	6	0.19	.6
	Experimental environment	14	0.25	.07				10	0.17	.4
	n-back version	14	0.06	.6				10	0.15	.3
	Modality of stimuli and responses	**14**	**0.22**	**<.0001**	-	-	-	**10**	**0.22**	**.001**
	Inter-stimulus interval	5	-0.03	.7	-	-	-	5	-0.09	.3
	Percentage of males	12	0.01	.07	-	-	-	9	0.01	.2
	Mean age	14	-0.002	.5	-	-	-	10	-0.002	.7
RR	Automation level	-	-	-	4	-0.02	.97	-	-	-
	Simulator fidelity	-	-	-	4	0.03	.97	-	-	-
	n-back version				4	-0.41	.052			
	Modality of stimuli and responses	-	-	-	4	0.03	.97	-	-	-
	Inter-stimulus interval	-	-	-	4	-0.1	.9	-	-	-
	Percentage of males	-	-	-	4	0.01	.6	-	-	-
	Mean age	-	-	-	4	0.01	.9	-	-	-

Notes: In this table and the following tables, n denotes the number of studies included; "-" means that there is no need for subgroup analysis, as the meta-analysis results did not demonstrate the presence of heterogeneity (see Table 4). The significant metrics (p<.05) are bold

Table 6. Results of sub-group moderator analyses.

Pairwise Comparison of Cognitive Load	Physiological Metrics	Moderator	Moderator Level	n	Pooled Z_r within Subgroup (95%CI)	Pooled r within Subgroup (95%CI)	Q_B	Q_W
L vs. M	**PS**	**n-back Version**	**Repeating**	**4**	0.66 (0.28, 1.03)	**0.58 (0.27, 0.77)**	17.36 (p<.001)	**10.83**
			Matching	**1**	0.53 (0.36,0.70)	**0.49 (0.35, 0.60)**		**-**
			Counting	**1**	-0.12 (-0.40,0.16)	-0.12 (-0.38, 0.16)		**-**
	EDA	Automation Level	L0	12	0.13 (0.06,0.19)	**0.13 (0.06, 0.19)**	33.82 (p<.001)	**4.10**
			L2	1	0.29 (0.05,0.53)	**0.29 (0.05, 0.53)**		-
			L3	1	1.10 (0.77,1.43)	**0.89 (0.65, 0.89)**		-
	EDA	n-back Modality	Auditory-verbal	11	0.12 (0.05,0.19)	**0.12 (0.05, 0.19)**	33.43 (p<.001)	**4.13**
			Auditory-manual	2	0.21 (0.07,0.36)	**0.21 (0.07, 0.36)**		**0.35**
			Visual-verbal	1	1.1 (0.77,1.43)	**0.89 (0.65, 0.89)**		-
L vs. H	HR	n-back Modality	Auditory-verbal	11	0.42 (0.25,0.58)	**0.41 (0.25, 0.56)**	27.42 (p<.001)	**43.31**
			Auditory-manual	4	0.21 (0.04, 0.38)	**0.21 (0.04, 0.36)**		**5.94**
			Visual-verbal	1	1.18 (0.86,1.50)	**0.84 (0.71, 0.91)**		**-**
	HR	Automation Level	L0	13	0.25 (0.19,0.32)	**0.25 (0.19, 0.31)**	57.97 (p<.001)	**12.03**
			L2	2	0.97 (0.73, 1.22)	**0.75 (0.66, 0.82)**		**1.38**
			L3	1	1.18 (0.86,1.50)	**0.84 (0.71, 0.91)**		**-**
	HR	Percentage of Male	<=50	9	0.31 (0.20, 0.42)	**0.30 (0.20, 0.40)**	9.85 (p=.007)	**17.73**
			(50-90)	3	0.49 (-0.36, 1.34)	0.46 (-0.32, 0.85)		32.54
			[90-100]	2	1.01 (0.58,1.43)	**0.77 (0.67, 0.86)**		**2.04**
	EDA		L0	8	0.08 (0.01,0.15)	**0.08 (0.01, 0.15)**		**6.18**

		Automation Level	L2	1	0 (-0.13,0.13)	0.00 (-0.13, 0.13)	38.08 (p<.001)	-
			L3	1	1.09 (0.76,1.42)	**0.8 (0.64,0.89)**		**-**
M vs. H	EDA	n-back Modality	Auditory-verbal	7	0.06 (-0.02,0.14)	0.06 (-0.02, 0.14)	7.60 (p<.001)	4.15
			Auditory-manual	2	0.21 (0.06,0.35)	**0.21 (0.06, 0.35)**		**0.18**
			visual-verbal	1	1.09 (0.76,1.42)	**0.8 (0.64,0.89)**		**-**

Notes: The bolded pooled r indicates significant (p<.05) associations; The bolded Q_W indicates the existence of heterogeneity of the metrics.

4. DISCUSSION

In this study, we systematically reviewed previous research regarding the associations between physiological and eye-tracking metrics and cognitive load levels in vehicles. The n-back tasks, which were commonly adopted to impose cognitive load in previous driving research, have been used as benchmarks of cognitive load levels (Janczewski et al., 2021). The random effects meta-analyses were conducted followed by moderator analyses.

4.1 Associations between the metrics and varying levels of cognitive load

Through meta-analyses, we found that although some metrics were found to be sensitive to the cognitive load in certain previous studies, they failed to pass the significance test in our meta-analyses. For example, the LF/HF ratio was found to be positively associated with the increase of cognitive load in Rahman et al (2020) and Zheng et al (2021), but it did not achieve statistical significance in our analyses.

At the same time, although it has been widely acknowledged that not all physiological features are sensitive to all levels of cognitive load (e.g., (Ayres et al., 2021; Li et al., 2022)) our meta-analyses provide further evidence to support this statement. For example, some metrics were more sensitive to lower levels of cognitive load compared to higher levels of cognitive load. We call these metrics **Sensitive-to-Low Metrics**. For example, the power of theta waves at Fp1 was sensitive to the low to median cognitive load levels with a median association (r=0.41), but it was not sensitive to higher levels of cognitive load (i.e., between medium and high cognitive load levels). This finding is partially in line with the findings in the meta-analysis by Chikhi et al (2022), who also observed a positive association between the power of θ wave at the frontal area and high cognitive load. But our research has provided higher resolution as we quantified more levels of cognitive load. Specifically, some metrics (e.g., θ-Fp1) may increase rapidly with a small increment of the cognitive load and reach a plateau at a medium level of cognitive load. This might be because some physiological indicators may cease to rise beyond a certain threshold, similar to some observations in brain studies (Bosking et al., 2017). The potential "ceiling effect" of θ-Fp1 may explain our observations, which has also been mentioned by Chikhi et al (2022) to explain the weaker association between the power of θ in multitasking versus single-tasking situation. It is also possible that beyond this "ceiling" point, other neural mechanisms or states can dominate (e.g., Weiss et al., 1995; Sauseng et al., 2004), and this may explain the significant difference between low to medium but insignificant difference between low to high cognitive load in terms of θ-Fp1. Though future research is needed to explain this phenomenon, the findings have some practical implications. Specifically, if we aim to detect medium (e.g., heavy traffic) to high cognitive load (e.g., heavy traffic and non-driving-related tasks) in drivers, the weights of sensitive-to-low metrics should be downgraded.

At the same time, some other metrics, for example, pupil size (PS), heart rate (HR), and skin conductance-electrodermal activity (SC-EDA) demonstrated a consistent growth relationship with the increase of cognitive load levels. We call these metrics

High-Resolution Metrics. Specifically, from low to medium, and from medium to high cognitive load, significant associations were observed for HR, SC-EDA, and PS. However, we should also be aware that, there are still differences in resolution among high-resolution metrics. Specifically, HR showed a substantially higher association strength with cognitive load variations than SC-EDA, and PS from eye-tracking measures exhibited an even higher association compared to HR and SC-EDA. The finding regarding HR is consistent with previous meta-analyses in non-driving domains (Hughes et al., 2019), which also found an association between high cognitive load and HR. In addition, the superior performance of PS is consistent with previous research (He et al., 2022) which found that, when predicting drivers' cognitive load states using five typical machine learning models, feature sets including eye-related measures could consistently result in high accuracies compared to feature sets with physiological measures alone. This highlights the superior performance of eye-related measures in monitoring drivers' cognitive load states (Chen et al., 2022).

In contrast to high-resolution metrics, some metrics were only sensitive from low to high cognitive load, but cannot differentiate low to medium and medium to high cognitive load levels. We call these **Low-Resolution Metrics**. For example, eye blink rate (EBR), respiration rate (RR), the total power (TP) demonstrated associations between low and high cognitive load levels only. This suggests that significant changes in cognitive load are required to induce notable variations in EBR, RR, and TP.

Finally, we also observed a non-linear relationship between the power spectrum of θ waves at Fp2 (θ-Fp2) and cognitive load levels. Specifically, we observed a positive correlation ($r = 0.41$) from low to medium cognitive load levels, but a negative correlation ($r = -0.24$) from medium to high cognitive load levels. Similar to the findings in θ-Fp1, additional states of the participants might have dominated the EEG at Fp2 when the task becomes "too difficult", which may have led to a decrease in θ-Fp2 at some point. In driver state monitoring systems, acknowledging this non-linearity is vital for accurate assessments.

It should be noted that the categorization of the metrics in this study is range-specific. Specifically, we only considered the cognitive levels from no secondary task to 2-back task. Even with no secondary task condition, drivers were still responsible for driving tasks. With lower or higher extreme cognitive load, high- or low-resolution metrics may be downgraded to sensitive-to-low ones; and with higher resolution of the cognitive task, the high-resolution metrics may become low-resolution ones. It should also be noted that previous research indicated that the driving performance measures may only be sensitive to high cognitive load (Yang et al., 2023). As for physiological and eye-tracking measures, we did not observe any that can differentiate medium to high cognitive load levels only. Thus, it seems that physiological and eye-tracking measures and driving performance measures may complement each other in driver-monitoring tasks.

4.2 Moderators

First of all, as expected, the n-back task version can moderate the association between pupil size and cognitive load from low to medium levels. Specifically, the repeating version of the n-back task led to significant associations in the repeating and matching versions, but not in the counting version. It is likely that the increased cognitive demand to remember the running total number of cases in the counting version led to an already high cognitive load even in the 1-back task and thus shadowed the effect of the additional cognitive load in the 2-back task. Though future research is needed to validate this hypothesis, the finding reveals the influence of task characteristics in modulating the cognitive-related eye-tracking metrics.

At the same time, we notice that the associations between HR and SC-EDA and cognitive load levels were moderated by the automation levels. Specifically, the associations between the HR and SC-EDA and the cognitive load were the strongest in vehicles with SAE Level 3 automation, both from low to medium and from low to high levels of cognitive load. It is possible that the drivers in SAL Level 3 vehicles are freed from continuous vehicle controlling tasks and they may experience the lowest workload in driving. Thus, the cognitive load imposed by the n-back task is less likely to be shadowed by the variations of task load in driving tasks. This finding suggests that different features and different algorithms may need to be designed for driver cognitive load detection in vehicles with driving automation, which is still lacking. To the best of our knowledge, only one study has focused on the driver cognitive load estimation algorithms in vehicles with driving automation (Meteier et al., 2021).

Moreover, in addition to automation levels, the associations were also moderated by the n-back task modality. Specifically, our analysis reveals that the correlation between HR and SC-EDA and cognitive load is most pronounced in the visual-verbal n-back tasks. It is likely that the high demand of visual resources in driving competes with the visual component in the visual-manual n-back tasks and thus leads to high sensitivity of the HR and SC-EDA to the visual-verbal n-back task. However, it should be noted that the visual component is not a cognitive component, and thus, the high associations of HR and SC-EDA might be the result of increased stress during the task (De Looff et al., 2018; Liu & Du, 2018) This finding indicates that the experiment settings may affect the physiological and eye-tracking measures (Nilsson et al., 2022).

Additionally, we observed that the characteristics of the participants may also affect the association between cognitive load and physiological responses, with male participants leading to a stronger association between HR and low to high levels of cognitive load. This highlights the importance of considering participant characteristics when designing driver monitoring systems.

Finally, it should be noted that heterogeneity has still been observed in most of the sub-groups, indicating that additional moderating factors may still exist. Future research is still needed to explore these factors and thus better guide the design of the driver monitoring systems.

4.3 Limitations and Future Directions

The current investigation presents several limitations. Firstly, the included investigations only considered the cognitive load imposed by the n-back tasks. Although n-back tasks have been widely adopted as a method for inducing cognitive load in traffic research. (Janczewski et al., 2021), other cognitive load induction tasks, such as mathematical tasks, may also be considered if their difficulty levels are quantified (Yang et al., 2023). Second, four studies were excluded from the present investigation due to the absence of required information for meta-analysis (Barua et al., 2017; Chihara et al., 2020; Solovey et al., 2014; Zhen et al., 2016). Consequently, though the meta-analysis based on a small sample size may still provide insights into potential trends and differences (Zheng, 2013), conclusions from our study should still be interpreted with caution, given that the associations between various metrics and cognitive load identified in our study are based on a limited number of studies. For similar reasons, the current investigation examined only a few potential moderators and some subgroups in the subgroup analyses contained relatively small numbers of studies. Finally, it should be noted that most of the research was conducted in simulators, and given the nature of the n-back task, some of the measures may be different from what they are in a natural driving condition (e.g., the response modality of the n-back task instead of cognitive load may have a strong influence on respiration rate and the complex lighting condition on a public road may shadow the influence of cognitive load on pupil size). Future research should be re-conducted when a larger sample size becomes available.

5. CONCLUSIONS

Despite the extensive research on the use of physiological and eye-tracking measures to assess cognitive load in driving, researchers have not reached a consensus on their associations with cognitive load. Based on a systematic review and a meta-analysis, for the first time, we quantified the association between physiological and eye-tracking metrics and cognitive load in driving. We identified four types of metrics, i.e., sensitive-to-low ones that can only differentiate the low (no secondary task or 0-back) to medium (1-back) level of cognitive load (including the power spectrum of θ waves of electroencephalogram at Fp1 channel); low-resolution ones that can only differentiate low and high cognitive load (including the overall power spectrum of electrocardiogram, eye blink rate and respiration rate) and others that show non-linear patterns with the increase of cognitive load (i.e., the power spectrum of θ waves at Fp2 channel). Furthermore, it has been found that n-back task versions, the modality of n-back tasks, the level of automation, and the percentage of male participants could moderate the associations between metrics and cognitive load.

This study, through a meta-analysis, offers a new perspective in understanding the relationship between physiological and eye-tracking metrics and different cognitive load levels and provides new insights into resolving the debates in this area. The findings highlight the importance of considering individual heterogeneity, driving automation, data collection environment, and metric characteristics when developing algorithms for driver cognitive load estimation. Future research should further validate our findings when more data and research become available.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 52202425), and in part by the Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), and Guangzhou Science and Technology Program City-University Joint Funding Project (No. 2023A03J0001).

CRediT AUTHORSHIP CONTRIBUTION STATEMNT

Ange Wang: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Chunxi Huang**: Validation, Formal analysis. **Jiyao Wang**: Data curation, Validation. **Dengbo He**: Formal analysis, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

DECLARATION OF COMPETING INTEREST

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

REFERENCES

Ayres, P. (2020). Something Old, Something New from Cognitive Load Theory. *Computers in Human Behavior*, *113*, 106503.

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load. *Frontiers in Psychology*, *12*, 702538.

Barua, S., Ahmed, M. U., & Begum, S. (2017, May). Classifying Drivers' Cognitive Load Using EEG Signals. *In pHealth* (pp. 99-106).

Bosking, W. H., Sun, P., Ozker, M., Pei, X., Foster, B. L., Beauchamp, M. S., & Yoshor, D. (2017). Saturation in Phosphene Size with Increasing Current Levels Delivered to Human Visual Cortex. *Journal of Neuroscience*, *37*(30), 7188–7197.

Broadbent, D. P., D'Innocenzo, G., Ellmers, T. J., Parsler, J., Szameitat, A. J., & Bishop, D. T. (2023). Cognitive load, Working Memory Capacity and Driving Performance: A Preliminary fNIRS and Eye Tracking Study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *92*, 121–132.

Cegovnik, T., Stojmenova, K., Jakus, G., & Sodnik, J. (2018). An Analysis of the Suitability of A Low-cost Eye Tracker for Assessing the Cognitive Load of Drivers. *Applied Ergonomics*, *68*, 1–11.

Chen, W., Sawaragi, T., & Hiraoka, T. (2022). Comparing Eye-tracking Metrics of Mental Workload Caused by NDRTs in Semi-autonomous Driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *89*, 109–128.

Chihara, T., Kobayashi, F., & Sakamoto, J. (2020). Evaluation of Mental Workload During Automobile Driving Using One-class Support Vector Machine with Eye Movement Data. *Applied Ergonomics*, *89*.

Chikhi, S., Matton, N., & Blanchet, S. (2022). EEG Power Spectral Measures of Cognitive Workload: A Meta-Analysis. Psychophysiology, 59(6), e14009.

Cohen, J. E. (1988). The Counterintuitive in Conflict and Cooperation. *American Scientist*, 76.6 (1988): 577-584.

Committee, O.-R. A. D. (ORAD). (2014). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. SAE International.

De Looff, P. C., Cornet, L. J. M., Embregts, P. J. C. M., Nijman, H. L. I., & Didden, H. C. M. (2018). Associations of Sympathetic and Parasympathetic Activity in Job Stress and Burnout: A Systematic Review. *Plos One*, *13*(10), e0205741.

Deng, M., Gluck, A., Zhao, Y., Li, D., Menassa, C. C., Kamat, V. R., & Brinkley, J. (2024). An Analysis of Physiological Responses as Indicators of Driver Takeover Readiness in Conditionally Automated Driving. *Accident Analysis & Prevention*, *195*, 107372.

Du, N., Kim, J., Zhou, F., Pulver, E., Tilbury, D. M., Robert, L. P., Pradhan, A. K., & Yang, X. J. (2020). Evaluating Effects of Cognitive Load, Takeover Request Lead Time, and Traffic Density on Drivers' Takeover Performance in Conditionally Automated Driving. *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 66–73.

Du, N., Yang, X. J., & Zhou, F. (2020). Psychophysiological Responses to Takeover Requests in Conditionally Automated Driving. *Accident Analysis and Prevention*,

Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Röning, J., Forlizzi, J. F., & Dey, A. K. (2014). Assessing Real-time Cognitive Load Based on Psycho-physiological

Measures for Younger and Older Adults. *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, 39–48.

Gable, T., Walker, B., Kun, A., Winton, R., & ACM. (2015). Comparing Heart Rate and Pupil Size as Objective Measures of Workload in the Driving Context: Initial Look. *In Adjunct proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications,* 20-25*,* New York, NY, USA.

Hajek, W., Gaponova, I., Fleischer, K. H., & Krems, J. (2013). Workload-adaptive Cruise Control – A New Generation of Advanced Driver Assistance Systems. *Transportation Research Part F: Traffic Psychology and Behaviour*, *20*, 108–120.

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An On-road Assessment of Cognitive Distraction: Impacts on Drivers' Visual Behavior and Braking Performance. *Accident Analysis & Prevention*, *39*(2), 372–379.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology*, 52, 139–183. North-Holland.

He, D., Donmez, B., Liu, C. C., & Plataniotis, K. N. (2019). High Cognitive Load Assessment in Drivers Through Wireless Electroencephalography and the Validation of a Modified N-Back Task. *IEEE Transactions on Human-Machine Systems*, *49*(4), 362–371.

He, D., Wang, Z., Khalil, E. B., Donmez, B., Qiao, G., & Kumar, S. (2022). Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transportation Research Record: Journal of the Transportation Research Board*, *2676*(10), 670–681.

Higgins, J. P. T., & Deeks, J. J. (2003). *Measuring Inconsistency in Meta-analyses | The BMJ*. https://www.bmj.com/content/327/7414/557.short.

Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac measures of cognitive workload: A meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 61(3), 393-414.

Jackson, L., Chapman, P., & Crundall, D. (2009). What Happens Next? Predicting Other Road Users' Behaviour as A Function of Driving Experience and Processing Time. *Ergonomics*, *52*(2), 154–164.

Li, P., Li, Y., Yao, Y., Wu, C., Nie, B., & Li, S. E. (2022). Sensitivity of Electrodermal Activity Features for Driver Arousal Measurement in Cognitive Load: The Application in Automated Driving Systems. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 14954–14967.

Lipsey, M., & Wilson, D. (2001). *Practical Meta-Analysis*. SAGE Publications.

Liu, Y., & Du, S. (2018). Psychological Stress Level Detection Based on Electrodermal Activity. *Behavioural Brain Research*, *341*, 50–53.

McCarthy, P. L., Holstein, S. A., Petrucci, M. T., Richardson, P. G., Hulin, C., Tosi, P., Bringhen, S., Musto, P., Anderson, K. C., Caillot, D., Gay, F., Moreau, P., Marit, G., Jung, S.-H., Yu, Z., Winograd, B., Knight, R. D., Palumbo, A., & Attal, M. (2017). Lenalidomide Maintenance After Autologous Stem-Cell Transplantation in Newly Diagnosed Multiple Myeloma: A Meta-Analysis. *Journal of Clinical Oncology*, *35*(29), 3279–3289.

Mehler, B., & Reimer, B. (2019). How Demanding is "Just Driving?" A Cognitive Workload—Psychophysiological Reference Evaluation. *Driving Assessment Conference*, *10*(2019), Article 2019.

Mehler, B., Reimer, B., & Dusek, J. A. (2011). *MIT AgeLab Delayed Digit Recall Task (N-Back)*. Cambridge, MA: Massachusetts Institute of Technology, 17.

Mehler, B., Reimer, B., & Coughlin, J. F. (2012b). Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task: An On-Road Study Across Three Age Groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(3), 396–412.

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*(1), 6–12.

Melnicuk, V., Thompson, S., Jennings, P., & Birrell, S. (2021). Effect of cognitive load on drivers' state and task performance during automated driving: Introducing a novel method for determining stabilisation time following take-over of control. *Accident Analysis & Prevention*, *151*, 105967.

Meteier, Q., Capallera, M., Ruffieux, S., Angelini, L., Abou Khaled, O., Mugellini, E., Widmer, M., & Sonderegger, A. (2021). Classification of Drivers' Workload Using Physiological Signals in Conditional Automation. *Frontiers in Psychology*, *12*, 596038.

Moher, D. (2009). *Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement | Annals of Internal Medicine*. https://www.acpjournals.org/doi/full/10.7326/0003-4819-151-4-200908180-00135

Muhrer, E., & Vollrath, M. (2011). The Effect of Visual and Cognitive Distraction on Driver's Anticipation in A Simulated Car Following Scenario. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(6), 555–566.

Muth, E. R., Moss, J. D., Rosopa, P. J., Salley, J. N., & Walker, A. D. (2012). Respiratory Sinus Arrhythmia as A Measure of Cognitive Workload. *International Journal of Psychophysiology*, *83*(1), 96–101.

Niezgoda, M., Tarnowski, A., Kruszewski, M., & Kamiński, T. (2015). Towards Testing Auditory-vocal Interfaces and Detecting Distraction While Driving: A

Comparison of Eye-movement Measures in the Assessment of Cognitive Workload. *Transportation Research Part F: Traffic Psychology and Behaviour*, *32*, 23–34.

Nilsson, E. J., Bärgman, J., Ljung Aust, M., Matthews, G., & Svanberg, B. (2022). Let Complexity Bring Clarity: A Multidimensional Assessment of Cognitive Load Using Physiological Measures. *Frontiers in Neuroergonomics*, *3*, 787295.

Rahman, H., Ahmed, M. U., Barua, S., & Begum, S. (2020). Non-contact-based Driver's Cognitive Load Classification Using Physiological and Vehicular Parameters. *Biomedical Signal Processing and Control*, *55*, 101634.

Recarte, M. A., & Nunes, L. M. (2000). Effects of Verbal and Spatial-imagery Tasks on Eye Fixations While Driving. *Journal of Experimental Psychology: Applied*, *6*(1), 31–43.

Reimer, B., Mehler, B., Coughlin, J. F., Godfrey, K. M., & Tan, C. (2009). An On-road Assessment of the Impact of Cognitive Workload on Physiological Arousal in Young Adult Drivers. *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 115–118, New York, NY, USA.

Rieck, J. R., DeSouza, B., Baracchini, G., & Grady, C. L. (2022). Reduced Modulation of BOLD Variability as A Function of Cognitive Load in Healthy Aging. *Neurobiology of Aging*, *112*, 215–230.

Sagberg, F., & Bjørnskau, T. (2006). Hazard Perception and Driving Experience Among Novice Drivers. *Accident Analysis & Prevention*, *38*(2), 407–414.

Sauseng, P., Klimesch, W., Doppelmayr, M., Hanslmayr, S., Schabus, M., & Gruber, W. R. (2004). Theta coupling in the human electroencephalogram during a working memory task. *Neuroscience Letters*, 354(2), 123-126.

Sârbescu, P., Stanojević, P., & Jovanović, D. (2014). A Cross-cultural Analysis of Aggressive Driving: Evidence from Serbia and Romania. *Transportation Research Part F: Traffic Psychology and Behaviour*, *24*, 210–217.

Schmidt, L., Shokraneh, F., Steinhausen, K., & Adams, C. E. (2019). Introducing RAPTOR: RevMan Parsing Tool for Reviewers | Systematic Reviews | Full Text. *Systematic Reviews*, *8*(1). https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-019-1070-0

Singh, S. (2015). Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. *Traffic Safety Facts - Crash Stats*, Article DOT HS 812 115. https://trid.trb.org/view.aspx?id=1346216&source=post_page

Solhjoo, S., Haigney, M. C., McBee, E., Van Merrienboer, J. J. G., Schuwirth, L., Artino, A. R., Battista, A., Ratcliffe, T. A., Lee, H. D., & Durning, S. J. (2019). Heart Rate and Heart Rate Variability Correlate with Clinical Reasoning Performance and Self-Reported Measures of Cognitive Load. *Scientific Reports*, *9*(1), 14668.

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014). Classifying Driver Workload Using Physiological and Driving Performance Data: Two Field Studies. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 4057–4066.

Spyridakos, P. D., Merat, N., Boer, E. R., & Markkula, G. M. (2020). Behavioural Validity of Driving Simulators for Prototype HMI Evaluation. *IET Intelligent Transport Systems*, *14*(6), 601–610.

Stapel, J., Mullakkal-Babu, F. A., & Happee, R. (2019). Automated Driving Reduces Perceived Workload, But Monitoring Causes Higher Cognitive Load Than Manual Driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *60*, 590–605.

Tjolleng, A., Jung, K., Hong, W., Lee, W., Lee, B., You, H., Son, J., & Park, S. (2017). Classification of A Driver's Cognitive Workload Levels Using Artificial Neural Network on ECG Signals. *Applied Ergonomics*, *59*, 326–332.

Viechtbauer, W. (2010). Conducting Meta-analyses in R with The Metafor Package. *Journal of Statistical Software*, *36*(3), 1–48.

von Janczewski, N., Wittmann, J., Engeln, A., Baumann, M., & Krauß, L. (2021). A Meta-analysis of The n-back Task While Driving and Its Effects on Cognitive Workload. *Transportation Research Part F: Traffic Psychology and Behaviour*, *76*, 269–285.

Weiss, T., Sust, M., Beyer, L., Hansen, E., Rost, R., & Schmalz, T. (1995). Theta Power Decreases in Preparation for Voluntary Isometric Contractions Performed with Maximal Subjective Effort. *Neuroscience Letters*, 193(3), 153-156.

Wickens, C. D. (2020). Processing Resources and Attention. *In Multiple Task Performance* (pp. 3-34). CRC Press.

Wickens, C. M., Mann, R. E., Stoduto, G., Ialomiteanu, A., & Smart, R. G. (2011). Age Group Differences in Self-reported Aggressive Driving Perpetration and Victimization. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(5), 400–412.

Yang, H., Liu, H., Hu, Z., Nguyen, A.-T., Guerra, T.-M., & Lv, C. (2024). Quantitative Identification of Driver Distraction: A Weakly Supervised Contrastive Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 25(2), 2034–2045.

Yang, H., Wu, J., Hu, Z., & Lv, C. (2024). Real-Time Driver Cognitive Workload Recognition: Attention-Enabled Learning with Multimodal Information Fusion. *IEEE Transactions on Industrial Electronics*, 71(5), 4999–5009.

Yang, S., Kuo, J., Lenne, M., Fitzharris, M., Horberry, T., Blay, K., Wood, D., Mulvihill, C., & Truche, C. (2021). The Impacts of Temporal Variation and

Individual Differences in Driver Cognitive Workload on ECG-Based Detection. *Human Factors*, *63*(5), 772–787.

Zhang, H., Qu, W., Ge, Y., Sun, X., & Zhang, K. (2017). Effect of Personality Traits, Age and Sex on Aggressive Driving: Psychometric Adaptation of the Driver Aggression Indicators Scale in China. *Accident Analysis & Prevention*, *103*, 29–36.

Zhang, Q., Yang, K., Qu, X., & Tao, D. (2022). Evaluation of Drivers' Mental Workload Based on Multi-modal Physiological Signals. *Shenzhen Daxue Xuebao (Ligong Ban)/Journal of Shenzhen University Science and Engineering*, *39*(3), 278–286.

Zheng, L., Qiao, X.-Q., Ni, T., Yang, W., & Li, Y.-N. (2021). Driver Cognitive Loads Based on Multi-dimensional Information Feature Analysis. *Zhongguo Gonglu Xuebao/China Journal of Highway and Transport*, *34*(4), 240–250.

Zhenhai, G., Yang, L., Lifei, D., Hui, Z., & Kaishu, Z. (2016). Driver Workload Evaluation Using Physiological Indices in Dual-Task Driving Conditions. *International Conference on Applied System Innovation*, 809–814, Osaka, Japan.

Zheng, M. (2013). *Application and Case Analysis of Meta-Analysis Software*. ISBN: 9787117171670, People's Health Publishing House, Beijing, China.