The Potential of Cognitive Circles to Measure Mental Load

Anonymous Author(s)

Abstract

In Human-Computer Interaction, Usability, and Interaction Design, obtaining objective measures of mental workload is desirable yet challenging, as current methods are either costly and intrusive or subjective and unreliable. To overcome these limitations, we devised Cognitive Circles, a technique that estimates workload by analyzing the kinematic properties of circular traces drawn on a tablet as people simultaneously perform cognitively demanding tasks of different types (arithmetic, reading, and spatial reasoning). We investigate the feasibility of this approach and lay the foundations for establishing its viability through a controlled experiment that addresses two questions: (A) Do participants' traces reliably encode information to predict the tasks' difficulty? and (B) Do predictive patterns generalize across tasks in different cognitive activities? Our results show that Cognitive Circles can predict task difficulty with accuracies reaching up to 94%, capturing meaningful signatures of mental workload (A). Prediction performance, however, varies substantially across task types (B), suggesting that each task domain induces people to exhibit distinct kinematic patterns. These findings highlight Cognitive Circles as a promising low-cost approach to workload assessment and point to its potential for informing adaptive HCI and the design of cognitively aware systems.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**.

Keywords

Mental Load Estimation, Mental Load Prediction, Cognitive Load

ACM Reference Format:

Anonymous Author(s). 2025. The Potential of Cognitive Circles to Measure Mental Load. In *Proceedings of User Interface Software and Technology (UIST '25).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Mental load is a fundamental construct in Cognitive Ergonomics and Human-computer Interaction (HCI). For instance, understanding the mental load associated with specific tasks and interfaces can help researchers compare alternative designs and, eventually, achieve more effective interfaces that do not overload users.

However, mental load is notoriously difficult to assess for three main reasons. First, it is a conceptually muddled construct, with variants such as mental workload, cognitive workload, and cognitive load—each meaning slightly different things across communities [46]. Second, mental load is most often measured by asking participants how easy or difficult a task was to complete or how difficult the interface was to use. This is achieved through *subjective assessments* that are noisy and can interrupt the task [31]. Third, existing objective measurement methods (e.g., EEG [25], oxygenation/blood flow detection [28]) tend to be expensive, intrusive, and cumbersome to use by designers and researchers.

In this paper, we propose Cognitive Circles, a technique for objectively measuring mental load without the cost and difficulty of other alternatives. The idea is simple: while users carry out a primary task, whose mental load we want to measure, they also trace circles on a tablet. The traces collected are processed and analyzed by machine learning algorithms able to differentiate between traces from trials with high and low mental load.

To determine the feasibility of the proposed technique, we must first address the following two questions: A) Do touch traces carried out during a task contain information about the difficulty of the task?; and B) Do traces from different types of tasks (e.g., arithmetic vs. language-based) share the same patterns that differentiate between easy and difficult instances of each task type? Answering the first question is important because it determines whether the main principle underpinning Cognitive Circles is supported by evidence. Answering the second question would tell us whether a generic dataset of circle traces labeled by difficulty is sufficient for general applications or whether task-specific datasets are required.

We carried out an experiment with 48 participants to answer the questions above. Participants completed easy and difficult tasks of three different types: arithmetic calculations, a reading/linguistic task, and logical sequence puzzles. The results are encouraging: the accuracy of differentiating difficult from easy tasks on a test set of 8 participants (48 tasks) was 75%, suggesting that the response to question A is positive. We also found that the model's performance varies across tasks (reaching 94% for sequential reasoning tasks) and that training the model on one task is as efficient as training the model on all tasks, suggesting that different tasks show different intrinsic patterns (question B). Our contributions are therefore:

- Cognitive Circles, a novel way to measure mental load and task difficulty;
- An experiment that validates the technique's potential and offers insights into how to develop it for testing different tasks.

While Cognitive Circles is still in its early stages, our findings lay the groundwork for its development as an experimental tool. They offer guidance on data collection, model training, and design considerations, moving us closer to a non-intrusive and inexpensive method for objective assessment of mental load and task difficulty.

2 Background and Related Work

There are many different terms to refer to concepts related to mental load. These differ in subtle ways, including the underlying assumptions, their theoretical underpinnings, and the researcher's

UIST '25, September 28–October 1, 2025, Busan, Korea

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

background. This multiplicity and subtlety often leads to misuse of the terms [22, 46] and, we believe, constitutes an obstacle to the effective use of the mental load as a conceptual and empirical tool.

The term *mental load* is preferred in the human factors domain, where it has a long history [40]. Most conceptualizations of mental load emerge from a systems/computational view of human cognition where the human is modeled as an information processor that has limited capacity and that experiences different levels of load as it carries out different tasks (e.g., [45]). In turn, different mental load levels result in different performances or outcomes of the processing [13, 33]. Performance and load are often associated in non-linear ways (e.g., performance may initially increase with load but then severely decline after capacity has been exceeded). Capacity is sometimes equated to memory at the lowest level of cognitive analysis [36], but it might include cognitive resources such as attention. Sometimes load is used to try to discern the cognitive architecture of the brain, such as whether there are different units that process different types of information or whether there are parallel or sequential processors [18].

Cognitive load is the term preferred in education, where the prominent *Cognitive Load Theory* [3] posits that learning tasks are only possible when there is sufficient human cognitive capacity to support the intrinsic load generated by the task, the extraneous load generated by the way in which the task is presented, and the task of learning itself (germane load) [21]. In this area, cognitive workload is usually defined more abstractly and, unfortunately, there is no consensus regarding how to separate its different components effectively [21], or about whether these three components are distinguishable or even explanatory [17].

Despite the difficulties navigating the conceptual landscape (some researchers consider this exercise moot [40]), it is clearly desirable to know whether a human is under heavy or light mental load, among other reasons, to be able to allocate tasks according to capacity dynamically (reducing the possibility of overload), to compare the difficulty of tasks, to compare the capacity of individuals, or to compare the benefits of using different tools or representations for carrying out the same task [22]. With these strong motivations, it is unsurprising that much effort has gone into studying and developing ways to measure cognitive or mental load or workload, although it is often not clear what is the correspondence between a measurement and the theoretical construct [40, 46].

In this paper, we purposefully remain agnostic about the theoretical background and use the term *mental load* because we find it suitably general and less associated with a specific theory. Also, we propose Cognitive Circles as a new way to measure mental load but do not yet posit or hypothesize the mechanisms or cognitive user models that underpin the phenomena captured by the technique.

2.1 Mental Load Measurements

Several surveys provide categorizations and characterizations of mental/cognitive load measurements (e.g., [4, 7, 22, 23, 34]). The most basic measurement is performance itself (usually completion time and error rates); however, these measurements provide limited insight into the reasons or origins of the load and do not differentiate mental load from other factors that could also affect performance.

Many surveys acknowledge subjective user questionnaires (e.g., SMEQ [51], NASA-TLX [10, 11]) as the most common method to ascertain load, but they also highlight their shortcomings, namely the need to interrupt the task or wait till its end to measure and the problems associated to it being self-reported (e.g., unstable, relative, prone to memory distortions). Alternative approaches, like the *instantaneous self-assessment* (ISA) [43], have attempted to address these concerns by reducing the time and cognitive overhead involved.

On the other hand, objective measurements include a large number of bodily signals, including cardiovascular (such as heart rate, ECG and heart rate variability [9]), brain-based (EEG [25], fMRI [41], fNIRS [28]), respiration [14], skin electrical and temperature changes [32], and eye-based [24] (gaze movement and pupil size). All these objective measurements require expensive, specialized sensors with different degrees of intrusiveness: from wearing a bracelet or a ring (e.g., to capture heart rate and galvanic skin response) to lying down in a constrained space (e.g., in a magnetic resonance imaging machine).

Another class of objective methods, more closely related to the approach we propose in this paper, is often referred to as *behavioral measurements*, which provide quantifiable data based on observable actions rather than physiological signals. For instance, research has tried to infer cognitive load from prosodic features of speech, such as pauses and speech rate [19, 44, 47]. Of particular interest to us are the works of Ruiz et al. [4, 35–37], who analyzed pen strokes in a range of tasks to infer the load that they impose, Luria and Rosenblum [27] and Yu et al. [49, 50], who focus on the handwriting signal, and Mock et al. [30], who analyze touch signals.

While behavioral methods offer a less intrusive alternative to physiological measurements, they are often highly tied to the nature of the task, making load inferences less reliable when user interaction is minimal (e.g., tasks involving few pen strokes). In addition, certain tasks demand more frequent or faster inputs, which may appear to reflect a higher mental load even when they do not. Our goal is to develop a measurement technique that is relatively insensitive to the content of the primary task, though not entirely task-agnostic (i.e., as explained in question B in the Introduction, different tasks such as reading, or arithmetic might require their own training). Ideally, by consistently performing the same secondary motor task (tracing circles on a tablet), we could obtain stable and comparable measurements across at least tasks with different contents. However, the trade-off of requiring a secondary task is that, while relatively unobtrusive, it may limit the technique's applicability in situations where two hands are needed for the primary task (see also Section 6.4 in the Discussion).

2.2 Mental Load and Motor Behavior

Prior research has consistently demonstrated that increased mental load can significantly affect motor performance through changes in kinematics and neuromuscular control (e.g., [20, 26, 39, 52]). This empirical evidence shows that when people are concurrently subject to cognitive and physical demands, they tend to simplify movement patterns, reduce motor precision and smoothness, increase movement variability, and allocate attentional resources differently, typically prioritizing one task at the expense of the other [1, 48].

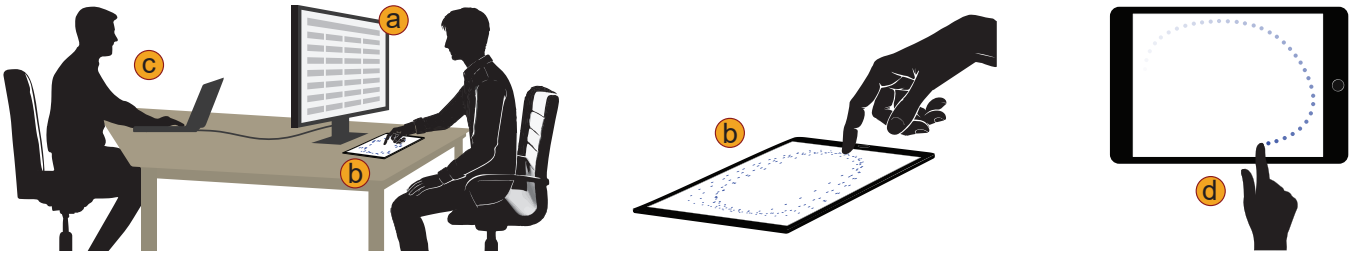


Figure 1: Experimental setup. Stimuli for the primary task appear on a monitor (a); participants trace circles on a tablet (b). The experimenter manages the experiment with a laptop (c). The circular traces become invisible after an initial period (d).

Herrebrøden et al. [12] found that solving arithmetic problems while using a rowing ergometer significantly reduced movement complexity, leading athletes—even elite rowers—to regress to simpler, less coordinated motor patterns. Similarly, DiDomenico and Nussbaum identified that the increase in various forms of physical activity generally resulted in diminished performance during arithmetic tasks [6].

The rationale behind Cognitive Circles builds upon these findings. We hypothesize that motor outputs during a continuous task (tracing circles) can serve as sensitive indicators of mental load.

3 Experiment

To assess the Cognitive Circles technique's potential to detect task difficulty and therefore mental load, we carried out a controlled empirical study. The main goal was to answer two questions: A) Do touch traces carried out during a task contain information about the difficulty of the task?; and B) Do traces from different types of tasks (e.g., arithmetic vs. language-based) share the same patterns that differentiate between easy and difficult performances?

3.1 Participants

We recruited participants from a local university in two phases. Group A consisted of 40 people (18 females, 22 males, 18–32 years old, mean age 22, median 21, 2 left-handed), and Group B consisted of 8 participants (3 females, 5 males, 21–33 years old, mean and median 26, 1 left-handed). Group B's data was exclusively used for testing, hence we did not use it for any training or model hyperparameter tuning. The demographic makeup, self-assessment of problem-solving and arithmetic skills, as well as the reading habits of both participant groups, are depicted in Figure 1 of the supplementary materials document.

3.2 Tasks, Apparatus and Procedure

The experiment's trials consisted of participants carrying out the secondary task, i.e., tracing circles on a tablet horizontally placed on the table, at the same time that they completed one of three possible primary tasks on an 18-inch vertical monitor in front (see Figure 1.a). Each trial lasted two minutes. For the secondary task participants were instructed to use their index finger of their dominant hand to continuously trace circles on a blank tablet, without specific instructions on the features of the circles (e.g., size, center, drawing direction). At the start of each trial, participants traced circles for five seconds, after which an auditory cue prompted them to begin the primary task on the main monitor (see Figure 1.a). At the beginning of the trial, touches left visible traces on the tablet; after 15 seconds they disappeared to avoid distraction.

We selected a set of three primary tasks based on three criteria: A) trials should represent a variety of different cognitive tasks involving different cognitive systems; B) it should be easy to design instances of the task that are easy and difficult, and; C) tasks should be familiar to participants. These criteria support a selection of tasks helpful to answer the two main questions of the study at this early stage of the development of the technique. The selected primary tasks were arithmetic, spatial reasoning, and reading.

Arithmetic task: Each trial consisted of 45 arithmetic comparisons (e.g., "3 + 6 is smaller than 1") displayed in a 9x5 grid. Participants had to count the number of true comparative statements within the set of 45 and report that number verbally at the end of the two minutes. The *easy arithmetic tasks* compared a small number against the outcome of an addition or a subtraction, whereas *difficult arithmetic tasks* compared two operations with larger numbers that could include multiplication and division (Figure 2(a)).

Spatial reasoning task: Each trial consisted of 45 graphical sequences of four figures, each similar to one of Raven's Progressive Matrices [2, 16]. Participants counted and verbally reported to the experimenter the number of sequences in which the fourth figure was logically consistent with the progression (i.e., correct). Easy tasks involved few elements and basic transformations (rotations or mirroring), whereas the difficult transformations featured more intricate objects with combinations of transformations (e.g., simultaneous mirroring and figure splitting—see Figure 2(b)).

Reading task: A brief short story[1] appeared on screen. In the easy version of the task participants had to count simple grammatical elements (e.g., "*How many punctuation marks are there in the following text?*"), whereas the difficult version required deeper linguistic analysis (e.g., "*How many masculine plural adjectives appear in the following text?*"—see Figure 2(c)).

After each trial, participants ranked the task difficulty on a 7-point Likert scale, and as a binary assessment (easy/difficult). The experiment lasted approximately 60 minutes. Participants received an introduction, signed consent, completed a demographic questionnaire and received an explanation of all tasks types. Then they familiarized themselves with the circle tracing task on the tablet (2 minutes). The bulk of the experiment consisted of carrying out

[1]The original text was presented to participants in Spanish, as they were all Spanish speakers. All experimental materials have been translated for this submission

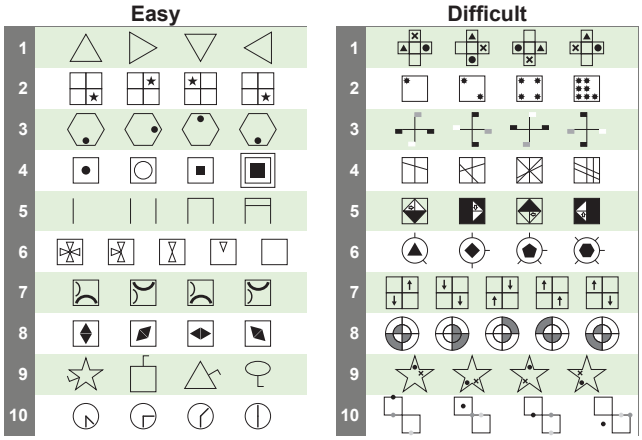
Easy

1	3+6	IS SMALLER THAN	1
2	5	IS GREATER THAN	2+4
3	1	IS GREATER THAN	6-2
4	9	IS GREATER THAN	7+8
5	30	IS GREATER THAN	23+5
6	29	IS SMALLER THAN	17+8
7	2	IS SMALLER THAN	3+9
8	3+8	IS GREATER THAN	4
9	2+5	IS GREATER THAN	6
10	3+9	IS SMALLER THAN	21

Difficult

1	43-36	IS GREATER THAN	12+7
2	42-7	IS GREATER THAN	31+3
3	23+19	IS SMALLER THAN	45-12
4	56-14	IS SMALLER THAN	15-23
5	34-6	IS SMALLER THAN	29+9
6	63+9	IS GREATER THAN	43-29
7	26x4	IS SMALLER THAN	45+57
8	46-13	IS GREATER THAN	7x4
9	18x3	IS SMALLER THAN	29+12
10	23+58	IS GREATER THAN	36x2

(a) Arithmetic



(b) Spatial reasoning

Easy

1	How many punctuation marks are there in the following text?
2	How many words in the following text contain the letter 'p' or 'P'?

Difficult

1	How many singular feminine nouns are there in the following text?
2	How many masculine plural adjectives are there in the following text?

Text (excerpt)

A story is a short narrative created by one or more authors, based on real or fictional events, whose plot is carried out by a small group of characters and with a simple plot. The story is shared both orally and in writing; Although at first, the most common was by oral tradition.

Furthermore, it can account for real or fantastic events, but always on the basis of being an act of fiction, or a mixture of fiction with real events and characters. It usually contains several characters who participate in a single central action, and there are those who believe that a shocking ending is an essential requirement of this genre. Its goal is to

(c) Reading

Figure 2: Task examples by type and difficulty level.

an easy and difficult trial of each of the three task types. Task type order was counterbalanced across participants and within each task type, the easy and difficult task appeared in any order. Participants were asked to prioritize correctness over speed, and reported outcomes at the end of the two minutes, regardless of whether they had completed all the trial's subtasks. They also had to verbally report how far they got in the list of subtasks: for arithmetic and spatial reasoning tasks this meant reporting the number in the subtask label, and for the reading task, reporting the last word they read.

3.3 Data Collection, Augmentation, and Pre-processing

We collected 4 main types of data: A) demographic information; B) participant subjective estimations of each trial's difficulty in two forms, a binary response (easy vs. difficult), and as a 7-point Likert scale; C) the participant response to the task (task counts and progress—see Section 3.2), and; D) the traces of the cognitive circles. For each participant, we computed their average Likert difficulty score and subtracted it from all their reported scores to account for individual rating differences. Our preliminary tests confirmed that this signal is less noisy than the raw Likert difficulty score and it is the one we used for our analyses.

Each trace is a bi-variate time series (a timestamped list of x and y coordinates). Because the sampling rate of touch points in a tablet is irregular, we resampled the original signals to achieve a uniform frequency of 66.6 Hz using linear interpolation.[2] The resampling ensured consistent temporal resolution while preserving the spatial characteristics of the original trace. From the x and y coordinates, we derived additional channels (features): linear velocity, linear acceleration, radius (distance of the point to a reference point or center), radial velocity, and angular velocity. The details of the derived features and the resampling are available in Section 2 of the supplementary materials. The final analyzed augmented cognitive circles data consisted of 7 channels with 7874 data points each.

4 Predictive Pipelines and Analysis

To answer our research questions, we first needed to find out whether it would be possible to predict the difficulty of the task using the trace data. Difficulty in our study can mean one of two things. The *designed difficulty* is one of the two levels of difficulty—easy or difficult—with which a trial was created (see Section 3.2). The *perceived difficulty* is an assessment that participants completed after each trial and had two forms: a binary (easy vs. difficult) judgment and a 7-point Likert-scale (see Section 3.3).

Our approach was to try a wide range of machine-learning techniques. The degree to which the best of the techniques predicts either the *designed difficulty* or the *perceived difficulty* provides us with a lower bound of how useful the participants' traces can be for workload prediction. While it is beyond the scope of this work to provide a comprehensive survey of all possible learning algorithms, we selected a diverse range of combinations of data representations and model types expected to perform well with this type of data.

[2]We used the `DataFrame.interpolate` https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html method from the Pandas Python library.

4.1 Data Representations

We selected data representations based on the literature and our experience designing, training, and evaluating learning pipelines.

Plain time series: This representation corresponds to the raw data consisting of seven channels: the original x and y coordinates and the five additional channels derived from them (linear velocity, linear acceleration, radius, radial velocity, and angular velocity), as described in Section 3.3.

Channel means representation: We can model a trace as a set of seven average values, one per channel in the plain time series. The main advantage of this representation is that the resulting features are interpretable (they correspond to averages of the traces' movement properties).

MiniROCKET features: We also used the output of applying the MiniROCKET algorithm [5], a state-of-the-art approach to learning feature maps, to the seven channels of the plain time series. MiniROCKET applies a set of random convolutional kernels to the data. The results, after pooling, provide a latent representation of the time series that can be sent as input to any state-of-the-art classifier or regressor. Contrary to the channel means representation, the MiniROCKET representation is not interpretable.

Trace Plots (VLAIR): These are 2D visual representations of the traces (bitmaps), generated from the time series. The visuals encode the touch points as constant-sized circles, and the circles are connected by straight lines colored on a gradient according to their instantaneous linear velocity (see Figure 3). Inspired by the VLAIR approach [15], this representation is meant to be accessible to humans and to computer vision models (see Section 4.2 below). The visual mapping was chosen iteratively, testing on a validation set taken from the training set (never on the testing set).

4.2 Models

The models that we can train are dependent on the representation and the labels that we predict. For the plain time series with binary labels we train an LSTM model with one layer, 20 (designed difficulty) or 10 (reported difficulty) hidden states, and dropout rates of 0.2 (designed) and 0.1 (reported). For the channel means and MiniROCKET feature representations we trained random forest classifiers with 100 tree estimators. For the trace plots (VLAIR), we employed a CNN classifier based on MobileNetV2 [38] augmented with two additional fully connected layers. All models were trained on the training set data (40 participants, 240 traces in total).

(a) Easy task (b) Difficult task (c) Linear Velocity

Figure 3: Traces made by a participant during the easy (a) and difficult (b) arithmetic tasks. Color encodes linear velocity normalized using min-max scaling to a $[0, 1]$ range (c).

To predict the Likert-reported difficulty, we used the channel means and MiniROCKET representations to train a random forest regressor and a pattern-based linear regressor using the HiPaR [8] algorithm. HiPaR first learns a baseline linear model on the entire dataset, whose predictions can be improved via a set of local linear models defined on subsets of the data. Those subsets are characterized by conditions on the input features, e.g., ($velocity > 0.5$ or $task = arithmetic$). Local models are learned only if they can help improve the performance of the baseline linear model. The details of the models and data representations, as well as the code, are in Section 4 of the Supplementary Materials.

5 Results

We used the pipelines described above to carry out multiple analyses that would allow us to address the two main questions of our study.

5.1 Do touch traces contain information about the primary task's difficulty?

The three leftmost numerical columns of Table 1 show a summary of the performance of different pairs of representation and model on the binary (easy vs. difficult) labels. The best-performing pipelines used the Random Forest (RF) classifier on the channel means and MiniROCKET representations, both with 0.75 accuracy. Performance is generally comparable across pipelines, with the

Table 1: Prediction performance of cognitive circles for the designed and perceived difficulty of the tasks in the test set (8-participant cohort). The table shows the results with and without including the task types as input feature.

Difficulty	Classifier	Cognitive Circles			Cognitive Circles + Task Type		
		Accuracy	F1 - difficult	F1 - easy	Accuracy	F1 - difficult	F1 - easy
Designed	RF on channel means	0.75	0.75	0.75	0.75	0.75	0.75
	RF on MiniROCKET	0.75	0.75	0.75	0.75	0.75	0.75
	LSTM on orig. series	0.75	0.75	0.75	0.75	0.75	0.75
	MobileNetV2 on raster images	0.71	0.73	0.68	0.73	0.78	0.67
Reported	RF on channel means	0.69	0.75	0.59	0.71	0.77	0.61
	RF on MiniROCKET	0.69	0.75	0.59	0.69	0.75	0.59
	LSTM on orig. series	0.65	0.70	0.56	0.69	0.72	0.65
	MobileNetV2 on raster images	0.63	0.75	0.25	0.63	0.73	0.40

Table 2: Ranking of features using the permutation feature importance score of the random forest classifiers trained on the channel means representation for predicting difficulty and task type.

Designed difficulty		Perceived difficulty		Task type	
Feature	Importance (std)	Feature	Importance (std)	Feature	Importance (std)
Ang. velocity	0.107 (0.098)	*Ang. velocity*	0.049 (0.050)	*Lin. velocity*	0.099 (0.032)
X coord	0.086 (0.035)	*Radial velocity*	0.024 (0.026)	*Radial velocity*	0.068 (0.037)
Radial velocity	0.079 (0.037)	*X coord*	0.017 (0.021)	*Ang. velocity*	0.061 (0.031)
Lin. velocity	0.066 (0.053)	*Lin. acceleration*	-0.007 (0.023)	*Y coord*	0.051 (0.031)
Radius	0.037 (0.039)	*Radius*	-0.009 (0.024)	*X coord*	0.028 (0.025)
Lin. acceleration	0.035 (0.038)	*Y coord*	-0.011 (0.030)	*Radius*	0.020 (0.030)
Y coord	0.033 (0.055)	*Lin. velocity*	-0.016 (0.032)	*Lin. acceleration*	0.008 (0.023)

VLAIR approach exhibiting a small disadvantage (likely due to its encoding of only three of the seven available channels—x, y, and *linear velocity*). In all prediction settings, accuracy was above 50%, which demonstrates that the traces do contain information about both the designed difficulty of the task and the perceived difficulty; however, all pipelines are better at predicting designed difficulty than perceived difficulty (see also Section 6).

The pipeline with random forests on the channel means allows us to measure the importance of the different features (channels) for the sake of interpretability. The two leftmost columns in Table 2 show the rank of features calculated using permutation importance (damage to prediction accuracy when noise is added to a feature by permuting values across rows). Angular velocity, the X coordinate and radial velocity are the top three features for predicting both designed and perceived difficulty (although in different orders).

Table 3: Performance predicting Likert reported difficulty using and omitting the task type as input feature (R^2).

Method	Without Task Type	With Task Type
RF regressor on MiniROCKET	0.11	0.29
RF regressor on Channel means	0.12	0.38
HiPaR on MiniROCKET	0.02	0.32
HiPaR on Channel means	-0.008	0.54

The results from the regressors, which predict the 7-point scale of perceived difficulty (left column of Table 3) show that the random forest regressor on the channel means and MiniROCKET representations could explain some of the variance (R^2 test scores of 0.12 and 0.11, respectively), whereas the HiPaR regressor's performance is almost negligible on both representations (R^2 of 0.02 on MiniROCKET and −0.008 on the channel means).

5.2 Do traces from different task types share patterns differentiating trial difficulty?

To answer this question we disaggregate the performance of the classifiers by task type (Table 4). Cognitive circles are best for predicting designed difficulty in spatial reasoning tasks and worst for reading tasks. The differences in accuracy between tasks are substantial and consistent across all classifiers and data representations. For the binary reported difficulty, the classifiers tend to perform worse at detecting easy tasks, more evidently in reading and spatial reasoning tasks due to class imbalance (e.g., 28 out of 40 spatial tasks were perceived as difficult), and despite taking the imbalance into account when training the classifiers. Learning individual difficulty classifiers on the traces of each task type led to lower accuracy for the random forest on the channel means pipeline and very similar performance for the random forest on MiniROCKET (the numbers between parentheses in the leftmost numerical column of Table 4).

In a further set of analyses, we compared the prediction performance of all pipelines with a version where an additional feature encoded which task a trial corresponds to. The results (in the rightmost columns of Table 1) indicate that including the task type does not significantly enhance the prediction of designed difficulty, and provides only minimal benefit for reported difficulty. The disaggregation of the performance of these task-aware classifiers (in Table 5) shows a very similar picture: explicitly marking the task

Table 4: Prediction performance of different classifiers when using the cognitive circles as input for predicting the designed and binary perceived difficulties on the three types of tasks. The accuracy scores in parentheses correspond to the performance of a classifier trained exclusively on the traces of that particular task type.

Difficulty	Task	RF on channel Means			RF on MiniROCKET			LSTM on orig. series			MobileNetV2 on images		
		Acc	F1 (d)	F1 (e)	Acc	F1 (d)	F1 (e)	Acc	F1 (d)	F1 (e)	Acc	F1 (d)	F1 (e)
Designed	Arithmetic	0.75 (0.63)	0.75	0.75	0.75 (0.75)	0.75	0.75	0.69	0.71	0.67	0.69	0.74	0.61
	Reading	0.63 (0.56)	0.63	0.63	0.63 (0.63)	0.63	0.63	0.69	0.67	0.71	0.50	0.50	0.50
	Spatial Rsn.	0.88 (0.88)	0.88	0.88	0.88 (0.88)	0.88	0.88	0.88	0.88	0.88	0.94	0.94	0.93
Reported	Arithmetic	0.63 (0.56)	0.57	0.67	0.63 (0.63)	0.57	0.67	0.63	0.57	0.67	0.50	0.64	0.20
	Reading	0.56 (0.44)	0.59	0.53	0.56 (0.50)	0.59	0.53	0.50	0.50	0.50	0.63	0.75	0.25
	Spatial Rsn.	0.88 (0.88)	0.93	0.50	0.88 (0.88)	0.93	0.50	0.81	0.89	0.40	0.75	0.85	0.33

Table 5: Prediction performance of different classifiers when using the cognitive circles and the task type as input for predicting the designed and binary perceived difficulties on the three types of tasks.

Difficulty	Task	RF on Channel Means			RF on MiniROCKET			LSTM on orig. series			MobileNetV2 on images		
		Acc	F1 (d)	F1 (e)	Acc	F1 (d)	F1 (e)	Acc	F1 (d)	F1 (e)	Acc	F1 (d)	F1 (e)
Designed	Arithmetic	0.75	0.75	0.75	0.75	0.75	0.75	0.69	0.71	0.67	0.93	0.74	0.62
	Reading	0.63	0.63	0.63	0.63	0.63	0.63	0.69	0.67	0.71	0.63	0.70	0.50
	Spatial Rsn.	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.89	0.86
Reported	Arithmetic	0.63	0.57	0.67	0.63	0.57	0.67	0.56	0.22	0.70	0.56	0.63	0.46
	Reading	0.56	0.59	0.53	0.56	0.59	0.53	0.56	0.53	0.59	0.56	0.67	0.37
	Spatial Rsn.	0.94	0.97	0.67	0.88	0.93	0.50	0.94	0.97	0.67	0.75	0.85	0.33

type in the data produces better results only for some tasks, e.g., an increase of 6 points for the random forests on channel means for the spatial reasoning task (compare Tables 4 and 5).

A similar analysis approach but on the Likert scale ratings shows that most regressors' accuracies are boosted when we encode task type in the data (see Table 3). The increase is most dramatic for the HiPaR pipeline on the channel means representation, which goes from an R^2 of -0.008 without task type information to 0.54.

We conducted a complementary analysis to predict the task type from the circle traces rather than its difficulty level. Table 6 shows the performance for the random forest approaches trained on the channels means and MiniROCKET representations, resulting in an overall accuracy of 0.75 (spatial reasoning tasks are predicted best—0.88) in both cases. For the random forest classifier trained on the channel means—the rightmost column of Table 2—we can see that the most informative channels for task prediction were linear velocity, radial velocity, and angular velocity.

Table 6: Performance of cognitive circles with a random forest for predicting the type of task. Numbers correspond to both channel means and MiniROCKET data representations.

F1 score			Accuracy
Arithmetic	Reading	Spatial reasoning	
0.63	0.75	0.88	0.75

6 Discussion

We first interpret the results to answer the two research questions. Then we discuss what the answers mean for the implementation of cognitive circles as a practical way to measure mental workload. Finally, we highlight the limitations of the study and the approach.

6.1 Touch traces do contain information about the task's difficulty (RQ-A)

The evidence from our experiment supports an affirmative answer to our first research question. All prediction pipelines achieved overall accuracies above chance (50% for a binary prediction), ranging from 71% to 75% when estimating the designed difficulty (i.e., the difficulty with which we created the tasks) and between 63% and 69% for the perceived difficulty (the binary difficulty label reported by participants). We found that predictions from the circle traces were more accurate for designed difficulty than for perceived difficulty, suggesting that the traces capture more information about the objective (criteria-based) difficulty of tasks than about participants' judgments. This finding lends some support to the validity of our approach and existing reliability concerns and inherent noise associated with subjective assessments.

Despite the affirmative answer, it is difficult to know, at this moment, whether the information carried by the traces is sufficient for reliably measuring workload in practical scenarios. On one hand, our experiments distinguished only between clearly defined *easy* and *difficult* tasks within a controlled environment, achieving accuracy levels potentially insufficient for real-world assessments (e.g., to compare two subtly different interfaces to complete a task). On the other hand, our accuracy results represent only a lower bound of what is achievable because: A) predictions were evaluated on data from participants that were never seen by the learning algorithms (i.e., all our tests were highly stringent); B) the number of participants is relatively small, and; C) our exploration of machine learning approaches was not fully comprehensive. Another promising indication from our analyses is that one regression model (HiPaR on channel means) reached an R^2 of 0.54 for the 7-point scale difficulty. We discuss promising options to further improve accuracy in Subsection 6.3.

6.2 Different tasks have shared indications of difficulty but significant differences (RQ-B)

The evidence from our experiment indicates that the answer to our second question is also affirmative: training on traces from multiple tasks improves predictive performance in other tasks. However, the magnitude of this improvement is modest, and variability in performance across tasks remains substantial. Furthermore, explicitly providing the task type as input improves prediction accuracy for some algorithms. Together with the fact that the task itself can be predicted from the traces (Table 6), this means that the patterns are fairly different between tasks.

Despite our efforts to balance the difficulty of trials across the different tasks, our experimental design does not allow us to determine whether these pattern differences arise from disparities in how we calibrated difficulty (an "easy" spatial reasoning task might be harder than a "difficult" reading task), or from intrinsic differences in the cognitive processes required by each task type. Nevertheless, the result has immediate implications for the future development of Cognitive Circles as a viable measuring technique.

To achieve sufficient accuracy in realistic scenarios, it will likely be necessary to train predictive models on traces covering a range of task types and a finer scale of difficulty levels. In retrospect, expecting a single machine learning model to generalize effectively across highly diverse cognitive tasks and capture a universal workload construct was overly optimistic.

6.3 Next Steps

The results from our analyses indicate several promising avenues to improve the accuracy and practical applicability of cognitive circles as a mental load measurement technique. A critical first step is to obtain larger datasets that reflect variability in human motor behavior in a more comprehensive way—potentially achievable through remote data collection. Additionally, expanding data collection to cover non-binary ranges of difficulties would enable continuous estimates of mental workload, significantly increasing the method's practical value for tasks requiring more subtle discriminations.

Future training datasets will also need to encompass a broader and more representative selection of tasks relevant to practitioners. Other tasks could include, for example, visual comparison, visual search, and different low-level subtasks of sensemaking. If a general model is not sufficiently accurate, explicitly providing the task type as an additional input to the machine learning pipeline might help increase predictive performance, as suggested by our results.

Finally, for more accurate and subtle measurements, a hybrid approach combining general population-based models with participant-specific calibration seems promising. In other words, if we ask participants to perform tasks with known, *a priori* difficulty levels, it would become feasible to derive personalized workload estimates of greater precision.

6.4 Limitations and Open Questions

An inherent limitation of the cognitive circles approach is that, in practice, individuals rarely perform secondary motor tasks with one hand while interacting with information. This limitation might be mitigated by employing the non-dominant hand or, when both hands are needed for the primary task, by tracking movements from other body parts, such as the feet [29, 42]. We also suspect that the degree of information contained in the traces might vary depending on the level of motor activity required by the primary task—our study examined only tasks without explicit motor requirements (e.g., moving the mouse, pressing keys).

There are several important open questions that will need to be addressed in future work. First, it is unclear whether increasing dataset size and ecological validity—through remote, varied data collection methods (e.g., online, with different devices and setups)—would enhance the robustness of the data. Second, it is uncertain whether the notion of a common workload construct across diverse task types will stand scrutiny once a broader set of tasks is tested; if not, we might have to constrain ourselves to workload comparisons within the same type of task. In any case, future work should validate cognitive circles against established subjective and objective measures of mental workload.

7 Conclusion

This paper introduced Cognitive Circles, a novel approach that estimates mental workload by leveraging continuous circle-tracing behavior. We conducted a dual-task experiment in which participants traced circles on a tablet while completing cognitively demanding tasks of three types (arithmetic, spatial reasoning, and reading). Our analyses indicate that the kinematic features of the traces can predict both the designed and perceived difficulty of these tasks with accuracies up to 94%, highlighting that motor coordination signals provide valuable cues about cognitive load. Yet, our findings also reveal significant performance variability across task types, pointing to the need for task-aware modeling in certain contexts.

Overall, these results offer a promising yet preliminary demonstration of how Cognitive Circles can deliver cost-effective workload assessments. Ultimately, we envision that the technique may inform the design of cognitively adaptive systems, enabling more nuanced and potentially real-time assessments of mental workload across a wide range of scenarios.

References

[1] Emad Al-Yahya, Helen Dawes, Lesley Smith, Andrea Dennis, Ken Howells, and Janet Cockburn. 2011. Cognitive motor interference while walking: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews* 35, 3 (2011), 715–728. https://doi.org/10.1016/j.neubiorev.2010.08.008

[2] Henry R. Burke. 1958. Raven's Progressive Matrices: A Review and Critical Evaluation. *The Journal of Genetic Psychology* 93, 2 (1958), 199–228. https://doi.org/10.1080/00221325.1958.10532420

[3] Paul Chandler and John Sweller. 1991. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction* 8, 4 (Dec. 1991), 293–332. https://doi.org/10.1207/s1532690xci0804_2 Publisher: Routledge _eprint: https://doi.org/10.1207/s1532690xci0804_2.

[4] Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M. Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. 2013. Multimodal behavior and interaction as indicators of cognitive load. *ACM Trans. Interact. Intell. Syst.* 2, 4 (Jan. 2013), 22:1–22:36. https://doi.org/10.1145/2395123.2395127

[5] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. 2021. MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 248–257. https://doi.org/10.1145/3447548.3467231

[6] Angela DiDomenico and Maury A. Nussbaum. 2011. Effects of different physical workload parameters on mental workload and performance. *International Journal of Industrial Ergonomics* 41, 3 (2011), 255–260. https://doi.org/10.1016/j.ergon.2011.01.008

[7] Jesús Díaz-García, Inmaculada González-Ponce, José Carlos Ponce-Bordón, Miguel Ángel López-Gajardo, Iván Ramírez-Bravo, Ana Rubio-Morales, and Tomás García-Calvo. 2022. Mental Load and Fatigue Assessment Instruments: A Systematic Review. *International Journal of Environmental Research and Public Health* 19, 1 (Jan. 2022), 419. https://doi.org/10.3390/ijerph19010419 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Luis Galárraga, Olivier Pelgrin, and Alexandre Termier. 2021. HiPaR: Hierarchical Pattern-Aided Regression. In *Advances in Knowledge Discovery and Data Mining*, Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty (Eds.). Springer International Publishing, Cham, 320–332.

[9] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psychophysiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 301–310. https://doi.org/10.1145/1864349.1864395

[10] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[11] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[12] Henrik Herrebrøden, Alexander Refsum Jensenius, Thomas Espeseth, Laura Bishop, and Jonna Katariina Vuoskoski. 2023. Cognitive load causes kinematic

changes in both elite and non-elite rowers. *Human Movement Science* 90 (2023), 103113. https://doi.org/10.1016/j.humov.2023.103113

[13] G. Robert J. Hockey. 2003. *Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks.* IOS Press.

[14] M. Sazzad Hussain, Rafael A. Calvo, and Fang Chen. 2014. Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion During Affective Interference. *Interacting with Computers* 26, 3 (May 2014), 256–268. https://doi.org/10.1093/iwc/iwt032

[15] Ai Jiang, Miguel A. Nacenta, Kasim Terzic, and Juan Ye. 2021. Visualization as Intermediate Representations (VLAIR) for Human Activity Recognition. In *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Atlanta, GA, USA) *(PervasiveHealth '20)*. Association for Computing Machinery, New York, NY, USA, 201–210. https://doi.org/10.1145/3421937.3422015

[16] John and Jean Raven. 2003. *Raven Progressive Matrices.* Springer US, Boston, MA, 223–237. https://doi.org/10.1007/978-1-4615-0153-4_11

[17] Slava Kalyuga. 2011. Cognitive load theory: How many types of load does it really need? *Educational psychology review* 23 (2011), 1–19.

[18] Beth Kerr. 1973. Processing demands during mental operations. *Memory & Cognition* 1, 4 (Dec. 1973), 401–412. https://doi.org/10.3758/BF03208899

[19] M. Asif Khawaja, Natalie Ruiz, and Fang Chen. 2007. Potential speech features for cognitive load measurement. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces (OZCHI '07)*. Association for Computing Machinery, New York, NY, USA, 57–60. https://doi.org/10.1145/1324892.1324902

[20] Hyung-Sik Kim, Mi-Hyun Choi, Jin-Seung Choi, Hyun-Joo Kim, Sang-Pyo Hong, Jae-Hoon Jun, Gye-Rae Tack, Boseong Kim, Byung-Chan Min, Dae-Woon Lim, and Soon-Cheol Chung. 2013. Driving Performance Changes of Middle-Aged Experienced Taxi Drivers Due to Distraction Tasks during Unexpected Situations. *Perceptual and Motor Skills* 117, 2 (2013), 411–426. https://doi.org/10.2466/22.25.PMS.117x23z6 PMID: 24611246.

[21] Paul A. Kirschner, Paul Ayres, and Paul Chandler. 2011. Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior* 27, 1 (2011), 99–105. https://doi.org/10.1016/j.chb.2010.06.025 Current Research Topics in Cognitive Load Theory.

[22] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* 55, 13s (July 2023), 283:1–283:39. https://doi.org/10.1145/3582272

[23] Arthur F. Kramer. 1991. Physiological metrics of mental workload: A review of recent progress. In *Multiple Task Performance*. CRC Press. Num Pages: 50.

[24] Krzysztof Krejtz, Andrew T. Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE* 13, 9 (Sept. 2018), e0203629. https://doi.org/10.1371/journal.pone.0203629 Publisher: Public Library of Science.

[25] Naveen Kumar and Jyoti Kumar. 2016. Measurement of Cognitive Load in HCI Systems Using EEG Power Spectrum: An Experimental Study. *Procedia Computer Science* 84 (2016), 70–78. https://doi.org/10.1016/j.procs.2016.04.068 Proceeding of the Seventh International Conference on Intelligent Human Computer Interaction (IHCI 2015).

[26] Raphaël Laurin and Lucie Finez. 2020. Working memory capacity does not always promote dual-task motor performance: The case of juggling in soccer. *Scandinavian Journal of Psychology* 61, 2 (2020), 168–176. https://doi.org/10.1111/sjop.12589

[27] Gil Luria and Sara Rosenblum. 2012. A computerized multidimensional measurement of mental workload via handwriting analysis. *Behavior Research Methods* 44, 2 (June 2012), 575–586. https://doi.org/10.3758/s13428-011-0159-8

[28] Serena Midha, Horia A. Maior, Max L. Wilson, and Sarah Sharples. 2021. Measuring Mental Workload Variations in Office Work Tasks using fNIRS. *International Journal of Human-Computer Studies* 147 (March 2021), 102580. https://doi.org/10.1016/j.ijhcs.2020.102580

[29] Bharti Mishra, Shashikanta Tarai, Vinod Ratre, and Arindam Bit. 2023. Processing of attentional and emotional stimuli depends on retrospective response of foot pressure: Conceptualizing neuron-cognitive distribution in human brain. *Computers in Biology and Medicine* 164 (2023), 107186. https://doi.org/10.1016/j.compbiomed.2023.107186

[30] Philipp Mock, Peter Gerjets, Maike Tibus, Ulrich Trautwein, Korbinian Möller, and Wolfgang Rosenstiel. 2016. Using touchscreen interaction data to predict cognitive workload. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 349–356. https://doi.org/10.1145/2993148.2993202

[31] Laura M Naismith, Jeffrey J H Cheung, Charlotte Ringsted, and Rodrigo B Cavalcanti. 2015. Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education* 49, 8 (2015), 805–814. https://doi.org/10.1111/medu.12732

[32] Nargess Nourbakhsh, Yang Wang, Fang Chen, and Rafael A. Calvo. 2012. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12)*. Association for Computing Machinery, New York, NY, USA, 420–423. https://doi.org/10.1145/2414536.2414602

[33] Robert D. O'DONNELL. 1986. Workload assessment methodology. *Cognitive processes and performance* (1986). https://cir.nii.ac.jp/crid/1573105975427191296 Publisher: Wiley.

[34] Raja Parasuraman and Ranjana Mehta. 2015. Neuroergonomic Methods for the Evaluation of Physical and Cognitive Work. In *Evaluation of Human Work* (4 ed.). CRC Press. Num Pages: 32.

[35] Natalie Ruiz, Qian Qian Feng, Ronnie Taib, Tara Handke, and Fang Chen. 2010. Cognitive skills learning: pen input patterns in computer-based athlete training. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/1891903.1891955

[36] Natalie Ruiz, Ronnie Taib, and Fang Chen. 2011. Freeform pen-input as evidence of cognitive load and expertise. In *Proceedings of the 13th international conference on multimodal interfaces (ICMI '11)*. Association for Computing Machinery, New York, NY, USA, 185–188. https://doi.org/10.1145/2070481.2070511

[37] Natalie Ruiz, Ronnie Taib, Yu (David) Shi, Eric Choi, and Fang Chen. 2007. Using pen input features as indices of cognitive load. In *Proceedings of the 9th international conference on Multimodal interfaces (ICMI '07)*. Association for Computing Machinery, New York, NY, USA, 315–318. https://doi.org/10.1145/1322192.1322246

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[39] Sabine Schaefer and David Scornaienchi. 2020. Table Tennis Experts Outperform Novices in a Demanding Cognitive-Motor Dual-Task Situation. *Journal of Motor Behavior* 52, 2 (2020), 204–213. https://doi.org/10.1080/00222895.2019.1602506

[40] Sarah Sharples and Ted Megaw. 2015. Definition and Measurement of Human Workload. In *Evaluation of Human Work* (4 ed.). CRC Press. Num Pages: 34.

[41] Minoo Sisakhti, Perminder S. Sachdev, and Seyed Amir Hossein Batouli. 2021. The Effect of Cognitive Load on the Retrieval of Long-Term Memory: An fMRI Study. *Frontiers in Human Neuroscience* 15 (Oct. 2021). https://doi.org/10.3389/fnhum.2021.700146 Publisher: Frontiers.

[42] Gabriella H. Small, Lindsey K. Molina, and Richard R. Neptune. 2023. The influence of altered foot placement and cognitive load on balance control during walking in healthy young adults. *Gait & Posture* 103 (2023), 37–43. https://doi.org/10.1016/j.gaitpost.2023.04.007

[43] Andrew J. Tattersall and Penelope S. Foord. 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics* 39, 5 (May 1996), 740–748. https://doi.org/10.1080/00140139608964495

[44] Maria Vukovic, Melissa Stolar, and Margaret Lech. 2021. Cognitive Load Estimation From Speech Commands to Simulated Aircraft. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1011–1022. https://doi.org/10.1109/TASLP.2021.3057492 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[45] Christopher D. Wickens. 2008. Multiple Resources and Mental Workload. *Human Factors* 50, 3 (June 2008), 449–455. https://doi.org/10.1518/001872008X288394 Publisher: SAGE Publications Inc.

[46] Max L Wilson. 2024. Mental Workload vs Cognitive Load vs everything else in HCI. In *Proceedings of the Cognitive Personal Informatics Workshop 2024*. https://medium.com/@cogpi/mental-workload-vs-cognitive-load-vs-everything-else-in-hci-575722d14572

[47] Bo Yin, Natalie Ruiz, Fang Chen, and M. Asif Khawaja. 2007. Automatic cognitive load detection from speech features. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces (OZCHI '07)*. Association for Computing Machinery, New York, NY, USA, 249–255. https://doi.org/10.1145/1324892.1324946

[48] Galit Yogev-Seligmann, Jeffrey M. Hausdorff, and Nir Giladi. 2008. The role of executive function and attention in gait. *Movement Disorders* 23, 3 (2008), 329–342. https://doi.org/10.1002/mds.21720

[49] Kun Yu, Julien Epps, and Fang Chen. 2011. Cognitive load evaluation of handwriting using stroke-level features. In *Proceedings of the 16th international conference on Intelligent user interfaces (IUI '11)*. Association for Computing Machinery, New York, NY, USA, 423–426. https://doi.org/10.1145/1943403.1943481

[50] Kun Yu, Julien Epps, and Fang Chen. 2013. Mental Workload Classification via Online Writing Features. In *2013 12th International Conference on Document Analysis and Recognition*. 1110–1114. https://doi.org/10.1109/ICDAR.2013.225 ISSN: 2379-2140.

[51] Ferdinand Rudolf Hendrikus Zijlstra and L Van Doorn. 1985. The construction of a scale to measure subjective effort. *Delft, Netherlands* 43, 1985 (1985), 124–139.

[52] Lisa A. Zukowski, Jaclyn E. Tennant, Gozde Iyigun, Carol A. Giuliani, and Prudence Plummer. 2021. Dual-tasking impacts gait, cognitive performance, and gaze behavior during walking in a real-world environment in older adult fallers and non-fallers. *Experimental Gerontology* 150 (2021), 111342. https://doi.org/10.1016/j.exger.2021.111342