Received: 17 June 2022 | Accepted: 20 February 2023

DOI: 10.1111/bjep.12595

INVITED ARTICLE

Digital ink and differentiated subjective ratings for cognitive load measurement in middle childhood

Kristin Altmeyer[1] | Michael Barz[2,3] | Luisa Lauer[4] | Markus Peschel[4] | Daniel Sonntag[2,3] | Roland Brünken[1] | Sarah Malone[1]

[1]Department of Education, Saarland University, Saarbrücken, Germany

[2]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

[3]Applied Artificial Intelligence, Oldenburg University, Oldenburg, Germany

[4]Department of Physics, Saarland University, Saarbrücken, Germany

Correspondence

Kristin Altmeyer, Department of Education, Saarland University, Campus A4 2, Saarbrücken 66123, Germany.
Email: kristin.altmeyer@uni-saarland.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: 01JD1811

Abstract

Background: New methods are constantly being developed to adapt cognitive load measurement to different contexts. However, research on middle childhood students' cognitive load measurement is rare. Research indicates that the three cognitive load dimensions (intrinsic, extraneous, and germane) can be measured well in adults and teenagers using differentiated subjective rating instruments. Moreover, digital ink recorded by smartpens could serve as an indicator for cognitive load in adults.

Aims: With the present research, we aimed at investigating the relation between subjective cognitive load ratings, velocity and pressure measures recorded with a smartpen, and performance in standardized sketching tasks in middle childhood students.

Sample: Thirty-six children (age 7–12) participated at the university's laboratory.

Methods: The children performed two standardized sketching tasks, each in two versions. The induced intrinsic cognitive load or the extraneous cognitive load was varied between the versions. Digital ink was recorded while the children drew with a smartpen on real paper and after each task, they were asked to report their perceived intrinsic and extraneous cognitive load using a newly developed 5-item scale.

Results: Results indicated that cognitive load ratings as well as velocity and pressure measures were substantially related to the induced cognitive load and to performance in both sketching tasks. However, cognitive load ratings and smartpen measures were not substantially related.

Conclusions: Both subjective rating and digital ink hold potential for cognitive load and performance measurement. However, it is questionable whether they measure the exact same constructs.

KEYWORDS
assessment, cognitive load measurement, extraneous load, intrinsic load, primary school, smartpen

INTRODUCTION

Attempts to measure cognitive load multidimensionally in middle childhood are rare, and there are no instruments available specifically for primary students. In this study, we investigated two methods for differentiated measurement of cognitive load in this age group regarding their sensitivity with respect to the manipulation of different cognitive load dimensions.

Cognitive load theory

Cognitive load theory (*CLT*) is concerned with the ease of information processing in working memory (Sweller et al., 2019). According to the CLT, during learning and problem solving, a good deal of a learner's limited working memory resources is committed to handling the complexity of the actual learning task, the design of the instructional materials and the generative learning process itself. Based on these distinguishable demands, the CLT separates three types of cognitive load: Intrinsic, Extraneous and Germane Cognitive Load (*ICL, ECL* and *GCL*, Sweller, 2010; Sweller et al., 1998).

The level of intrinsic cognitive load (ICL) is assumed to be determined by the complexity of a current task and mainly by the number of elements that must be processed concurrently (Sweller et al., 2011). As the learning process itself reduces the number of interacting elements (Sweller et al., 2019), ICL also depends on the learner's prior knowledge.

Extraneous cognitive load is assumed to be elevated if cognitive resources are devoted to processes resulting from an inappropriate design of the learning materials. For instance, irrelevant processing takes place if related pieces of information are presented temporarily or spatially separated from each other and have to be integrated mentally (split attention effect; Cierniak et al., 2009; Schroeder & Cenkci, 2018). In contrast, if the learning content is presented in a way that facilitates the identification and integration of essential information, ECL is low. Sweller et al. (1998) assume that ECL should be reduced to free working memory resources, which is expected to result in more efficient learning.

Germane cognitive load is related to the working memory resources which are devoted to schema building. Therefore, high GCL is desirable as it is conducive to performance (Sweller, 2010). The most prominent CLT approach (e.g., Sweller et al., 1998) assumed that ICL, ECL and GCL are independent types of working memory load that add up to the total cognitive load. Recently, GCL was understood as a germane resource which can be available to the learner for dealing with the interactivity of the elements in the learning materials if it is not wasted on cognitive processes caused by a poor instructional design (Kalyuga, 2011; Sweller, 2010; Sweller et al., 2011, 2019).

Cognitive load measurement in children

Measuring cognitive load is essential for refining and proving the assumptions of the CLT as well as to develop load-optimizing strategies (Paas et al., 2003). Nevertheless, a valid measurement of cognitive load in children is considered challenging, and this challenge becomes even greater when aiming at capturing its multidimensional nature (Ayres & Paas, 2012; Kirschner et al., 2011).

Subjective cognitive load measurement

Concurrently, self-reports are the most prevalent approach to cognitive load measurement in the field of education (Anmarkrud et al., 2019). Usually, learners rate cognitive load perceived during a previously completed unit of instruction on a Likert scale referring to a single (e.g., Park et al., 2015; Stebner et al., 2017; Yung & Paas, 2015) or two subsequent items, often differentiating between mental effort and perceived difficulty (e.g., Eitel et al., 2014; Korbach et al., 2017; Lee & Mayer, 2015).

A few rating scales provide the learners with multiple items on the three cognitive load dimensions. Klepsch et al. (2017) and Klepsch and Seufert (2020) have carried out extensive validation studies of their differentiated cognitive load questionnaire, reporting two experimental studies which served to develop and adjust the items (Klepsch et al., 2017) and six further studies (Klepsch & Seufert, 2020) that proved the validity of the newly developed questionnaire for different types of tasks in a variety of learning domains. Internal consistency, criterion validity and prognostic validity were empirically supported for an instrument consisting of two items measuring ICL, three for ECL and two (plus one, if applicable) to measure GCL. Krieglstein et al. (2022) conducted a meta-analysis of studies that had applied differentiated cognitive load scales (Eysink et al., 2009; Klepsch et al., 2017; Leppink et al., 2013, 2014). Overall, the meta-analysis indicated that the scales were reliable and valid in measuring adult ICL, ECL and GCL.

Despite their obvious benefits, subjective measures also come with some disadvantages. In general, retrospective ratings can be biased by memory effects (Schmeck et al., 2015) and singular measurements cannot capture dynamic variations of cognitive load (Chen et al., 2016).

Beyond these general constraints, there are further limitations of subjective cognitive load measurement when referring to children as a target group. First, it is unclear whether children can reflect on their own learning processes. Evidence that children might be capable of retrospective metacognition regarding their learning processes and outcomes (Metcalfe & Finn, 2013) and producing reliable cognitive load ratings (Ayres, 2006) is met with concerns implying that particularly younger children might not be able to provide valid and reliable self-reports in general (Chambers & Johnston, 2002) and on their cognitive load specifically (Leahy, 2018).They might tend to give extreme ratings (Chambers & Craig, 1998) and have struggles understanding complex or negatively worded items (Marsh, 1986).

In their meta-analysis, Krieglstein et al. (2022) found no evidence that the multidimensional cognitive load scales were less appropriate for children than for adults. However, only a few studies with younger school children were included in the meta-analysis and the study with the youngest sample was the study by Tang et al. (2019), in which the participants had an average age of 11.1 years.

Accordingly, differentiating measurements of cognitive load can be expected to place high demands on self-reflection and text comprehension in children, which makes it necessary to approach their development with consideration. Some principles can be learned from the differentiated measurement of adults. For example, it can be recommended to only include items on load types that are varied in the experimental conditions (Korbach et al., 2019). Furthermore, in their study comparing two differentiated cognitive load measurements, Skulmowski and Rey (2020) found evidence that the choice of measurement instrument had a significant impact on cognitive load ratings and that the instrument should match the type of learning task. In this regard, the authors consider the cognitive load questionnaire by Klepsch et al. (2017) particularly well suited for interactive tasks.

Objective cognitive load measurement with smartpens

Objective measurement methods are based on observations and recordings of participants' behaviour or reactions that reflect their concurrent total cognitive load. Only in very few cases have objective measures been used to differentiate the three cognitive load dimensions (e.g., for eye tracking; Zu et al., 2020).

Objective cognitive load measurement methods include task performance (post-test scores, performance in primary or secondary tasks; e.g., Korbach et al., 2017), physiological (heart rate and variability, brain activity, galvanic skin response or pupil dilation; for an overview, see Chen et al., 2016) and behavioural load indicators. The latter are based on activities tracked during task solving, of which most can be observed in real time and non-intrusively. In prior research, for example, gestures, eye movements or computer mouse interaction were identified as load indicators (Chen et al., 2016).

A further activity that is assumed to be particularly closely linked to cognitive processes and performance in written assignments is writing and sketching behaviour. Both are goal directed activities involving complex interactions between the brain, hand and eyes. Since writing takes years to evolve, the writing process itself can be a challenge for children (Berninger, 1991). Therefore, sketching behaviour may be a less confounded cognitive load indicator in children.

Smartpens are used to record and analyse a set of dynamic handwriting signals. Smartpens that write on paper can be considered as particularly unobtrusive input devices, as they can hardly be distinguished from a usual pen and fit seamlessly into a child's familiar learning environment.

There are various theoretical approaches to identify smartpen measures on which cognitive load can have an impact. Since high working memory load is assumed to interfere with cognitive and psycho-motor tasks, high cognitive load is expected to affect completion time in written tasks (Ruiz et al., 2007). With regard to smartpen measures, this leads to the conclusion that cognitive load influences temporal measures like writing velocity. Easy tasks inducing low cognitive load might be solved faster. High cognitive load is furthermore associated with a high (written) energy expenditure which can be reflected in strong writing pressure (Oviatt et al., 2018, 2021). Additionally, it is assumed that low mental load is associated with automatic, consistent, fluent and smooth writing, reflected by low variability (i.e., standard deviation) in temporal and pressure-based smartpen measures (Luria & Rosenblum, 2012; Smits-Engelsman & Van Galen, 1997). In contrast, high cognitive load is assumed to lead to a dis-automatization causing an increase in variability of writing features (Van Gemmert & Van Galen, 1998).

These theoretical assumptions on predictive smartpen measures have also been supported in studies on computer mouse interaction as a related type of goal-driven fine motor activity performed by hand. For example, high load tasks were found to lead to a reduction of velocity of mouse cursor movements (Rheem et al., 2018) as well as to an increase in mouse click pressure (Witte et al., 2021).

Moreover, the supposed impact of cognitive load on smartpen measures has been confirmed in various studies on text writing, digit writing and sketching. Regarding text writing, several investigations used a sentence making task (Ransdell & Levy, 1999) to induce different mental workload levels. Yu et al. (2011) showed that writing pressure and writing velocity information were cognitive load indicators. Lin et al. (2013) revealed that a subset of writing features including average pressure, azimuth, velocity in Y-direction, count of sensible pen tip pauses and maximum pressure achieved a cross-validation accuracy of 76.27% for cognitive load level classification. Wu et al. (2016) confirmed that it is also possible to predict mental workload levels based on handwriting patterns for a target group of children. Badarna et al. (2018) demonstrated that measures of motor control as well as pressure and velocity were affected by the induced cognitive load level in a text writing task.

In a study on digit writing (Luria & Rosenblum, 2012), participants performed numerical progressions of varying difficulty. Results revealed that velocity handwriting measures were influenced by cognitive load. Moreover, the results indicated that high cognitive load was reflected in increased variation and dis-automatization.

In a subsequent study on sketching behaviour, Rosenblum and Luria (2016) compared complex figure drawing from memory or by copying with paragraph copying. They found mean pressure and mean velocity to indicate cognitive load variation. It was further demonstrated that the difficulty levels of standardized sketching tasks could be predicted with over 90% accuracy for a sample of primary school children by a relatively small number of parameters recorded by a smartpen (Barz et al., 2020).

This study

In this study, we aimed at developing and preliminarily validating a subjective instrument with the help of which younger children can rate their perceived ICL and ECL. The second aim was to examine the potential of different parameters of smartpen use during sketching to serve as real-time indicators of ICL and ECL.

For this purpose, we identified two standardized child-oriented tasks that, firstly, allowed accurate performance measurement (criterion), secondly, could be varied in terms of the level of ICL and ECL they generated and thirdly, required continuous working with a pen. We had primary school children complete these tasks using smartpens and presented them in high load and low load versions in a within-subjects design. Cognitive load was rated on a self-developed differentiated scale.

If the measurement instruments were sensitive to cognitive load, there should be both high criterion-based validity and high convergent validity of the measures. The following hypotheses were therefore formulated:

Reflection of cognitive load variation in subjective ratings:

1a. The children rate their perceived ICL higher in the high-ICL than in the low-ICL tasks.

1b. The children rate their perceived ECL higher in the high-ECL than in the low-ECL tasks.

Relationship between cognitive load ratings and performance:

2a. Children's ICL ratings are negatively related to their performance in the sketching tasks.

2b. Children's ECL ratings are negatively related to their performance in the sketching tasks.

Predictive value of smartpen measures for performance:

3a. Low mean pressure and high velocity as well as low standard deviations of pressure and velocity can predict high performance in the Trail Making Test for children (TMT-C).

3b. Low pressure and high velocity as well as low standard deviations of pressure and velocity can predict high performance in Drawing Patterns Tasks.

Differences in smartpen measures depending on cognitive load variation:

4a. While velocity is lower, pressure as well as standard deviations of pressure and velocity are higher in high-ICL than in low-ICL tasks.

4b. While velocity is lower, pressure as well as standard deviations of pressure and velocity are higher in high-ECL than in low-ECL tasks.

Predictive value of smartpen measures for cognitive load ratings:

5a. High mean pressure and low mean velocity as well as high standard deviations of pressure and velocity predict high ICL ratings.

5b. High mean pressure and low mean velocity as well as high standard deviations of pressure and velocity predict high ECL ratings.

METHOD

Sample

Data collection took place at a large university event for children, to which several hundred local children and their parents were invited. Thirty-six of these children voluntarily participated in this study. They were not expected to have any specific prior knowledge. Since seven smartpens were available at the same time, the children were placed in ad hoc grouping arrangements of five to seven children but did not interact as a group. The children that took part at the same time did not belong to the same class or school. Moreover, the individual children were placed at a distance from each other and worked quietly on their own. The data of three children had to be excluded (one was significantly older than the others, there were problems with data recording for another and a third did not understand the instructions for the cognitive load measurement). The 33 remaining children (56% female) were between 7 and 12 years old (M = 9.96; SD = 1.13).

Design and procedure

In a within-subjects design, all participants engaged with the material in all conditions. After the children and their parents had given their informed consent, the children were provided with a smartpen and the paper-pencil-based materials. All steps of the procedure were introduced verbally by two female experimenters according to a fixed script. For each measurement, there was an example exercise that the children solved simultaneously.

In step one, the children filled out a demographic questionnaire, then they completed the Trail Making Test for children (TMT-C; Reitan, 1992) two times either first in version A (low ICL) or first in version B (high ICL). After each trial, they filled out a cognitive load questionnaire on their perceived ICL and ECL. Subsequently, the children performed two versions of a subset of six items from the Drawing Patterns subtest of the Snijders-Oomen nonverbal intelligence test (Snijders et al., 2005). There were also two parallel versions of this test, A (low ECL) and B (high ECL), which children worked on one after the other in either order. After each part, again, they completed the cognitive load questionnaire. The whole procedure took about 30 min.

Measures

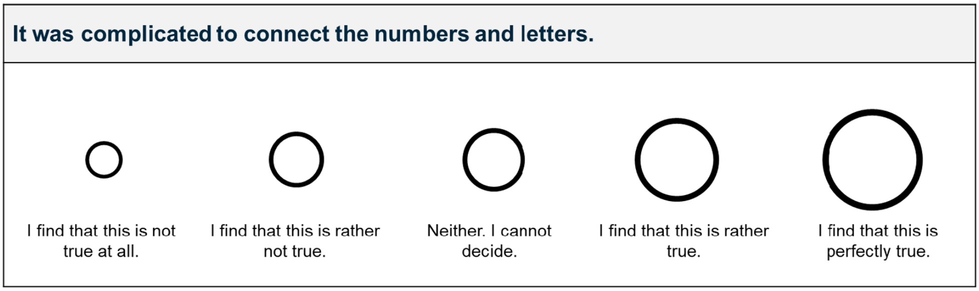
Cognitive load questionnaire

We developed a differentiated cognitive load rating questionnaire containing two items on ICL and three on ECL. Each item consisted of a statement (e.g., for ICL: 'The patterns that I had to draw were complicated') and a Likert scale on which the students indicated their agreement. The items were derived from the naïve rating scale of Klepsch et al. (2017) but adapted to the target group through various measures (see Table 1): Firstly, the items for GCL were excluded. Secondly, we simplified the wording and adapted it exactly to the respective sketching tasks. Third, a 5-point instead of the original 7-point Likert scale was used. Each level of agreement was described verbally (from 'I find that this is not true at all' to 'I

TABLE 1 Items of the differentiated measurement of intrinsic (ICL) and extraneous (ECL) cognitive load.

Type	Task	Item adapted to children	Original item (Klepsch et al., 2017)
ICL	TMT	While connecting the numbers (and letters), I had to keep many rules in mind at the same time.	For this task, many things needed to be kept in mind simultaneously.
	DP	When drawing the patterns, I had to follow many rules at the same time.	
ICL	TMT	Connecting the numbers (and letters) was complicated.	This task was very complex.
	DP	The patterns I had to draw were complicated.	
ECL	TMT	Connecting the numbers (and Letters) was difficult because important things were hidden.	During this task, it was exhausting to find the important information.
	DP	It was difficult to draw the patterns because important things were hidden.	
ECL	TMT	The way the task appeared on the sheet made it difficult to connect the numbers (and letters).	The design of this task was very inconvenient for learning.
	DP	The way the task appeared on the sheet made it difficult to draw the patterns.	
ECL	TMT	Connecting the numbers (and Letters) was difficult because important things were not visible.	During this task, it was difficult to recognize and link the crucial information
	DP	Drawing the patterns was difficult because important things were not always visible.	

Note: DP, Drawing Patterns sub-task of SON (Snijders et al., 2005); TMT, Trail Making Test (Reitan, 1992).



It was complicated to connect the numbers and letters.

○	○	○	○	○
I find that this is not true at all.	I find that this is rather not true.	Neither. I cannot decide.	I find that this is rather true.	I find that this is perfectly true.

FIGURE 1 Sample Item of the differentiated cognitive load questionnaire (ICL). *Note*: This sample item was presented after completion of the TMT-C.

find that this is perfectly true') and increasing levels of agreement were visualized with circles increasing in size (see Figure 1). Moreover, we developed a pre-instruction for children on how to assess and report one's own experiences with the help of the developed Likert scale. The supervisors explained that there were no correct or false answers and what the size of the circles meant. It was pointed out that children should express their very own experience and that they can take their time with the answers. After the instruction, open questions were clarified, the supervisors asked the children to complete the practice item ('It is very noisy in this classroom') and checked immediately for each child whether they had given a plausible answer. If necessary, they would have offered help to single children. However, it turned out that all children had marked plausible levels of agreement to the sample item and thus obviously had well understood how the items were to be completed. The main points of this pre-instruction were briefly repeated prior to each cognitive load rating questionnaire.

Trail Making Test for children (TMT-C)

The TMT-C (Reitan, 1992) has two versions: in version A, circled numbers from 1 to 15 are irregularly distributed on a sheet (Figure 2, Version A). The child's task is to connect the numbers in the correct order as quickly as possible. Version B contains a task switching component, as again 15 circled elements are distributed on the sheet, but these contain the numbers from 1 to 8 and the letters from A to H (Figure 2, Version B) that must be connected alternately (e.g., 1-A-2-B-3-…). We assumed version B to be more complex and to trigger higher ICL. Analysing the element activity of the task as suggested by Sweller (2010), higher ICL in version B is caused by altering the task material through adding sequential letters as additional elements that interact with the number elements and have to be processed simultaneously. Consequently, version A represents the low load condition and version B the high load condition.

As dependent variable, children's completion times for versions A and B were recorded by means of the smartpen.

Drawing Patterns Task

To vary ECL, we selected the Drawing Patterns subtest from the Snijders-Oomen nonverbal intelligence test (*SON*; Snijders et al., 2005), which requires patterns of increasing complexity to be drawn. In the original version (A, low ECL), each item consists of a pattern drawn of one or two black lines, with a small piece in the middle having been left out (see Figure 3a). Children are instructed to fill in the missing piece of the pattern. We created our own version B (high ECL) of six selected items (extracted from each of the two available parallel tests). In version B, the children were not allowed to fill in the missing piece of the pattern directly into the gap, but had to turn the sheet over and draw the missing piece on the reverse side. In terms of Sweller's (2010) suggested analysis of element interactivity, this leads to a

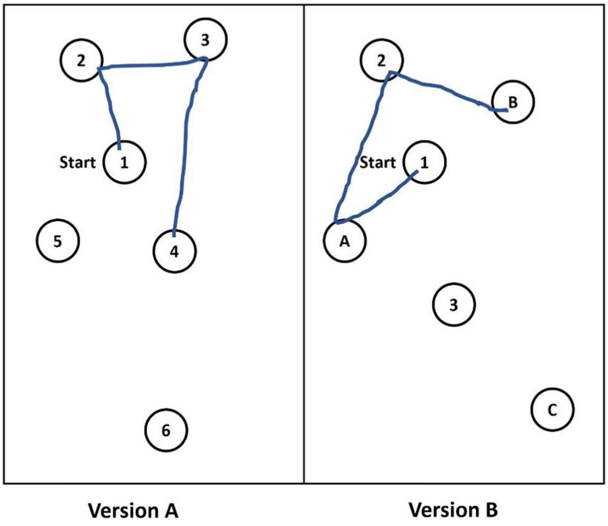


FIGURE 2 Fictitious excerpts from TMT-C Versions A (low load) and B (high load).

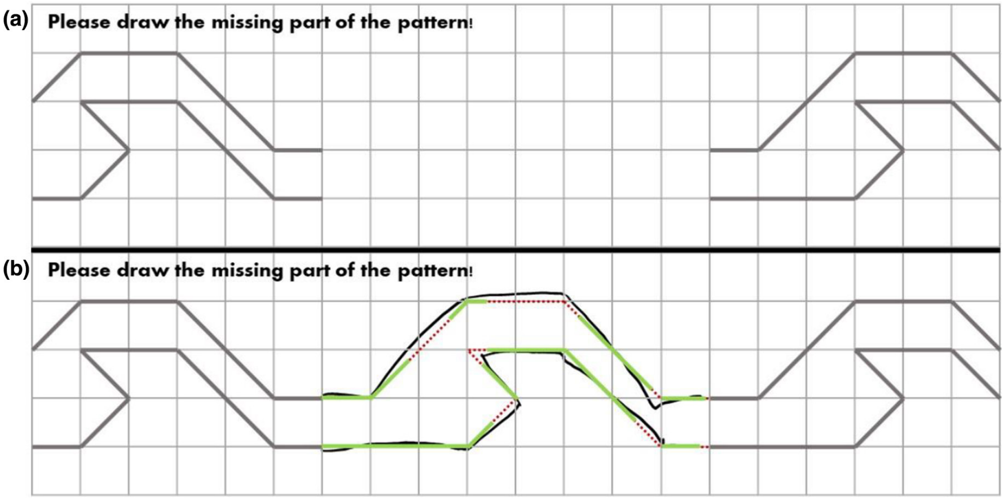

(a)
Please draw the missing part of the pattern!
(b)
Please draw the missing part of the pattern!

FIGURE 3 Fictitious sample item of the Drawing Patterns Task. *Note*: (a) Item as presented to the children; (b) completed item (black line), with ideal line for coverage calculation (green: tolerated parts with no or small deviation, red dotted: non-tolerated parts with large deviation from ideal line).

challenging search process in order to link the interacting elements reference pattern, empty grid and self-drawn lines. The children had to find the beginning and end points of the lines forming a pattern by repeatedly checking the reference pattern on the reverse side. This discontinuity was supposed to lead to a higher working memory load since the reference pattern had to be held in working memory while drawing. It was alternated between the participants which parallel test items were given in the versions A or B. According to the administration manual, scoring is done in a binary way using masks. Since the children drew with a smart pen, it was convenient to carry out a more precise scoring: we calculated the coverage of the ideal reference patterns by children's drawings. Slight deviations to the reference pattern were tolerated, as can be seen in Figure 3b. This was done by an automatic assessment algorithm based on the digital drawing, the reference pattern and a distance threshold for deviations. The resulting performance measure coverage describes the percentage to which a child's drawing matches the reference pattern within the tolerated deviation threshold. For more information on the described coverage metric, see also (Barz et al., 2020).

Digital ink

In our experiment, the children worked with a Neo Smartpen M1, which resembles an ordinary ballpoint pen but is equipped with a camera aligned with the writing surface and multiple pressure and motion sensors. To enable the recording of digital ink, all materials were printed on paper, which was previously imprinted with a special micro-dot pattern (NCode) and was used to locate the pen on the sheet by means of the camera. The children participating in this study were not aware of the smartpen's recordings: they experienced the sketching tasks as working with an ordinary pen on ordinary paper as they are used to from school. For more information on apparatus, see (Barz et al., 2020).

Smartpens record three basic features: *x* and *y* coordinates, pressure on the pen tip and time stamps. By means of these, a large set of sophisticated metrics of pen use can be calculated (for an overview, see Prange & Sonntag, 2022). For our purpose, only selected indicators for velocity and pressure of the normalized sketches were considered that had already yielded promising results in previous studies regarding cognitive load measurement. The extracted digital ink measures were based on the feature set by Willems and Niels (2008) and are listed and explained in Table 2.

RESULTS

A brief preregistration and data that support the findings of this study are available on the Open Science Framework: https://osf.io/p3e6d/.

In addition to conventional inferential analyses resulting in *p*-values, we performed bootstrapping with 5000 samples and provide the bias-corrected and accelerated bootstrap 90% confidence intervals (BCa 90% CI) for each analysis. Although bootstrapping might be more robust with respect to the statistical limitations of the present small sample (n = 33), it is nevertheless important to mention that the small sample size leads to a lack of power regarding the following analyses and that results should only be interpreted as indications in certain directions.

The within-subject variation in this study was achieved by providing the participants with high load and low load versions of tasks. While paired t-tests were used to compare variables between conditions, that is, high and low load versions of tasks, standard linear multiple regressions were carried out separately for each task version. Consequently, each multiple regression refers to task specific load ratings or performance as a criterion and smartpen measures that were precisely related to this task as predictors.

TABLE 2 Description of selected digital ink features.

Features	Type	Description
Average pressure	Pressure based	Force sensors built into the pen measure how hard the user presses the tip onto the surface. Average pressure is calculated for a set of sample points.
Standard deviation of pressure	Pressure based	Standard deviation of pressure is calculated for a set of sample points.
Average velocity	Temporal	Based on recorded time stamps for sample points (e.g., numbers in the TMT-C): Average velocity between sample points is calculated.
Standard deviation of velocity	Temporal	Standard deviation of the average velocity between sample points is calculated.

Note: The measures were computed according to the formulas provided with the feature set published by Willems and Niels (2008). All measures refer to recordings of sample points, each including location, pressure and time data.

TABLE 3 Means (*M*) and standard deviations (*SD*) for performance measures.

		Version	
		A **(*Low load*)**	***B*** **(*High load*)**
Completion time TMT-C	*M* (*SD*)	29.3 (12.93)	63.51 (25.33)
Coverage Drawing Patterns	*M* (*SD*)	77.83 (17.4)	61.1 (18.1)

Note: Completion time in seconds, coverage in percentage.

TABLE 4 Means (*M*) and standard deviations (*SD*) for subjective cognitive load ratings.

		TMT-C		Drawing Patterns	
Type of load		***A*** **(*Low load*)**	***B*** **(*High load*)**	***A*** **(*Low load*)**	***B*** **(*High load*)**
ICL	*M* (*SD*)	1.44 (.49)	2.06 (.92)	2.41 (.94)	3.38 (.8)
ECL	*M* (*SD*)	1.45 (.59)	1.93 (.95)	2.27 (1.13)	3.2 (1.04)

Abbreviations: ECL, Extraneous cognitive load; ICL, Intrinsic cognitive load.

Manipulation checks were performed to ascertain whether the B versions of both tasks resulted in worse performance than the A versions. *T*-tests for dependent samples indicated that students solved version A of the TMT faster on average than version B, ($t(32) = -9.26$, $p < .001$, BCa 90% CI [−40.76, −28.26] $d_z = -1.61$) and that they drew more percent of the missing patterns correctly in the A version of the Drawing Patterns than in the B version, ($t(32) = 4.89$, $p < .001$, BCa 90% CI [.11, .22], $d_z = .85$). Descriptive data on performance are displayed in Table 3. Accordingly, the data indicate that the manipulations were successful.

Hypothesis 1a. Cognitive load variation and ICL ratings

A dependent samples *t*-test demonstrated that students rated their ICL significantly different for the two versions of the TMT-C ($t(32) = -3.75$, $p < .001$, BCa 90% CI [−.88, −.39], $d_z = -.65$). The descriptive data indicate that they reported lower ICL for version A than for B (see Table 4). The two versions also evoked significantly different ECL ratings, ($t(32) = -2.59$, $p = .007$, BCa 90% CI [−.78, −.19], $d_z = -.45$). ECL was also rated lower for version A.

Hypothesis 1b. Cognitive load variation and ECL ratings

A dependent samples *t*-test demonstrated that the students' ECL ratings differed significantly for the two versions of the Drawing Patterns Task, ($t(32) = -5.01$, $p < .001$, BCa 90% CI [−1.28, −.61] $d_z = -.87$), indicating that higher ECL was experienced in version B (see Table 3). Moreover, the two versions of the Drawing Patterns Task differed regarding ICL ratings, ($t(32) = -6.44$, $p < .001$, BCa 90% CI [−1.21, −.74], $d_z = -1.12$). ICL was also rated higher for version B.

Hypothesis 2a. Performance and ICL ratings

Linear regressions revealed that ICL ratings related to TMT-C version A did not predict completion time for TMT-C version A ($F(1, 31) = .05$, $p = .412$, BCa 90% CI [−9.24, 5.98], $R^2_{adj} = -.03$), but ICL ratings for TMT-C version B predicted completion time for version B ($F(1, 31) = 18.65$, $p < .001$, BCa 90% CI [5.43, 24.45], $R^2_{adj} = .36$): The higher ICL was rated, the more time was needed for task completion. Moreover, high ICL ratings were related to worse performance in Drawing Patterns Task A ($F(1, 31) = 3.94$, $p = .028$, BCa 90% CI [−.12, −.001], $R^2_{adj} = .08$), but ICL did not predict performance in Task B (F(1, 31) = .04, BCa 90% CI [−.06, .08], $p = .427$, $R^2_{adj} = -.03$).

Hypothesis 2b. Performance and ECL ratings

Linear regressions revealed that ECL ratings predicted completion time for TMT-C version A ($F(1, 31) = 4.18$, $p = .025$, BCa 90% CI [.54, 16.09], $R^2_{adj} = .09$) and version B ($F(1, 31) = 17.51$, $p < .001$, BCa 90% CI [4.92, 22.53], $R^2_{adj} = .34$): The higher ECL was rated, the higher was a participant's task completion time. Furthermore, ECL ratings were negatively related to performance in Drawing Patterns Task A ($F(1, 31) = 11.51$, $p = .001$, BCa 90% CI [−.13, −.02], $R^2_{adj} = .25$), but did not predict performance in Task B ($F(1, 31) = 1.58$, $p = .109$, BCa 90% CI [−.09, .01], $R^2_{adj} = .02$).

Hypothesis 3a. Smartpen measures and performance in TMT-C

Descriptive data on smartpen measures can be found in Table 5.

Smartpen measures were recorded separately for each task version. A multiple linear regression including the smartpen measures means and standard deviations of pressure and velocity related to TMT-C version A as predictors for the dependent variable completion time in TMT-C version A was significant ($F(4, 28) = 13.43$, $p < .001$, $R^2_{adj} = .61$). Mean and standard deviation of velocity were significant predictors for performance in TMT-C version A (see Table 6 on regression coefficients and BCa 90% CIs resulting from bootstrapping). Higher mean velocity and less variation in velocity lead to better performance.

A multiple linear regression including the same smartpen measures but related to TMT-C version B as predictors and considering completion time in TMT-C version B as criterion was also signif-

TABLE 5 Means (*M*) and standard deviations (*SD*) for smartpen measures.

		TMT-C		**Drawing Patterns**	
		A (Low load)	***B (High load)***	***A (Low load)***	***B (High Load)***
Mean velocity	*M*(*SD*)	32.76 (9.1)	24.70 (6.12)	78.1 (81.5)	68.2 (15.8)
Velocity variation	*M*(*SD*)	31.35 (9.48)	25.89 (7.98)	52.3 (17.3)	53.6 (15.6)
Mean pressure	*M*(*SD*)	157.99 (37.63)	154.3 (35.8)	157.34 (35.49)	158.37 (34.91)
Pressure variation	*M*(*SD*)	34.16 (14.72)	37.94 (14.10)	36.2 (11.16)	37.52 (10.66)

Note: Velocity in smartpen units per second, pressure as raw Neo Smartpen M1 sensor output.

TABLE 6 Standardized regression coefficients and BCa 90% CIs of predictors of completion time in TMT-C Versions A and B.

	ß	***t*(28)**	***p***	**BCa 90% CI**
TMT-C Version A				
Mean velocity	−1.13	−4.52	<.001**	[−2.42, −1.12]
Velocity variation	.48	1.84	.038*	[.006, 1.54]
Mean pressure	−.07	−.6	.278	[−107.42, 59.31]
Pressure variation	−.21	1.58	.063	[−19.89, 331.11]
TMT-C Version B				
Mean velocity	−.58	−2.3	.015*	[−4.91, .34]
Velocity variation	.08	.3	.382	[−1.67, 1.99]
Mean pressure	−.37	−1.89	.07	[−589.56, 109.72]
Pressure variation	.26	1.29	.103	[−143.15, 985.16]

$^*p < .05$; $^{**}p < .001$.

icant (F(4, 28) = 2.97, p =.018, R^2_{adj} =.2). With respect to p-values of coefficients, mean velocity showed to significantly predict performance (see Table 6). Higher mean velocity was related with better performance in TMT-C version B. In contrast, since all coefficient BCa 90% CIs include zero, bootstrapping-based results revealed no significant smartpen predictors for TMT-C version B (see Table 6).

Hypothesis 3b. Smartpen measures and performance in Drawing Patterns Task

For the dependent variable performance in Drawing Patterns Task A, a multiple regression analysis including means and standard deviations of velocity and pressure related to Drawing Patterns Task A as predictors was not significant (F(4, 28) = .3, p =.438, R^2_{adj} = −.1).

With regard to the criterion performance in Drawing Patterns Task B, a multiple regression with according predictors but related to Drawing Patterns Task B was significant (F(4, 28) = 3.74, p =.007, R^2_{adj} =.26). Both mean and standard deviation of velocity predicted task performance (see Table 7 for regression coefficients and BCa 90% CIs resulting from bootstrapping). Higher mean velocity and less velocity variation were related with higher performance.

Hypothesis 4a. ICL variation and smartpen measures

Paired t-tests revealed that mean velocity was significantly higher in TMT-C version A than version B (t(32) = 5.96, p <.001, BCa 90% CI [.59, 10.42], d_z =1.04). Mean pressure did not differ between TMT-C versions (t(32) = 1.43, p =.81 BCa 90% CI [−1.55, 8.93], d_z =.25). Standard deviation of velocity was higher for TMT-C version A than B (t(32) = 4.19, p <.001, BCa 90% CI [3.25, 7.81], d_z =.729). In contrast, standard deviation for pressure was higher in TMT-C version B than in version A (t(32) = −2.99, p =.003, BCa 90% CI [−6.36, −1.2], d_z = −.52).

Hypothesis 4b. ECL Variation and smartpen measures

Paired t-tests showed no significant differences between Drawing Patterns Task A and B regarding mean velocity (t(32) = .76, p =.228, BCa 90% CI [−4.3, 33.8], d_z =.13) or mean pressure (t(32) = −.65, p =.261, BCa 90% CI [−3.81, 1.64], d_z = −.11). Moreover, there were no differences for standard deviation of velocity (t(32) = −.61, p =.274, BCa 90% CI [−4.3, 1.9], d_z = −.11) and pressure (t(32) = −1.5, p =.072, BCa 90% CI [−2.8, .11], d_z = −.26).

TABLE 7 Standardized regression coefficients and BCa 90% CIs of predictors of coverage regarding Drawing Patterns Task B.

	β	t(28)	p	BCa 90% CI
Drawing patterns Task B				
Mean velocity	1.18	3.21	.002*	[.62, 1.9]
Velocity variation	−1	−2.85	.004*	[−1.98, −.21]
Mean pressure	.26	1	.16	[−.001, .005]
Pressure variation	−.29	−1.06	.15	[−.01, .002]

*p <.05.

Hypothesis 5a. Smartpen measures and ICL rating

A multiple linear regression using means and standard deviations of velocity and pressure related to TMT-C version A as predictors for ICL ratings after TMT-C version A was not significant ($F(4, 28) = .69$, $p = .302$, $R^2_{adj} = -.04$). The corresponding multiple regression for the smartpen measures related to TMT-C version B and the criterion ICL ratings after TMT-C version B showed no significance either ($F(4, 28) = 1.75$, $p = .083$, $R^2_{adj} = .09$).

With regard to Drawing Patterns Task, a multiple regression analysis using smartpen features related to Drawing Patterns Task A as predictors were not significant for the dependent variable ICL ratings after Drawing Patterns Task A ($F(4, 28) = 1.06$, $p = .199$, $R^2_{adj} = .01$). Another multiple regression analysis considering smartpen measures related to Drawing Patterns Task B and ICL rating after Drawing Patterns Task B was not significant either ($F(4, 28) = .97$, $p = .219$, $R^2_{adj} = -.003$).

Hypotheses 5b. Smartpen measures and ECL rating

A multiple linear regression including means and standard deviations of velocity and pressure related to TMT-C version A as predictors for ECL ratings after TMT-C version A was not significant ($F(4, 28) = .12$, $p = .488$, $R^2_{adj} = -.12$). With regard to ECL ratings after TMT-C version B, a multiple regression using the smartpen measures related to TMT-C version B was significant ($F(4, 28) = 2.46$, $p = .034$, $R^2_{adj} = .16$). High mean velocity was related to low ECL ratings (see Table 8 for regression coefficients and BCa 90% CIs resulting from bootstrapping). A further multiple regression with smartpen measures related to Drawing Patterns Task A predicting the ECL rating after Drawing Patterns Task A showed no significance ($F(4, 28) = .43$, $p = .393$, $R^2_{adj} = -.08$). The multiple regression investigating the predictive power of the same smartpen measures but related to Drawing Patterns Task B for the criterion ECL ratings after Drawing Patterns Task B was also not significant ($F(4, 28) = 1.38$, $p = .133$, $R^2_{adj} = .05$).

DISCUSSION

This study investigated whether a newly developed subjective rating instrument and digital ink recorded by a smartpen were sensitive to manipulation of ICL and ECL in primary school students. Overall, cognitive load ratings as well as pressure and especially velocity-based smartpen measures were related to induced cognitive load and performance. However, cognitive load ratings and smartpen measures were not substantially related.

In line with hypotheses 1a and b, cognitive load variation was reflected by differences in children's ratings. In both sketching tasks, cognitive load was rated significantly lower in the conditions that were expected to induce low cognitive load than in the conditions that were expected to induce high cognitive load.

TABLE 8 Standardized regression coefficients and BCa 90% CIs of predictors of extraneous cognitive load rating after TMT-C Version B.

	β	$t(28)$	p	BCa 90% CI
TMT-C Version B				
Mean velocity	−.73	−2.82	.004*	[−.19, −.04]
Velocity variation	.27	1.02	.158	[−.01, .078]
Mean pressure	−.21	−1.06	.149	[−.02, .01]
Pressure variation	.05	.23	.409	[−.02, .03]

*$p < .05$.

Concerning hypothesis 2a and b, the results on the relation between sketching performance and cognitive load ratings were mixed. Highly rated ICL was associated with lower performance in TMT-C B as well as Drawing Patterns Task A. High-ECL was associated with low performance in TMT-C A and B and Drawing Patterns Task A. However, neither ICL nor ECL ratings were related to performance in the high-ECL inducing Drawing Patterns Task B. These results support the assumption of Chambers and Johnston (2002) that children might struggle with valid self-assessment, but the findings also indicate tha it depends on the specific task.

Confirming hypotheses 3a and b, smartpen measures held predictive value for task performance. High mean velocity proved to predict high task performance in TMT-C version A as well as in Drawing Patterns Task B. This is in line with prior results on writing velocity and mental workload (e.g., Yu et al., 2011). Moreover, low variation in velocity was associated with high performance for TMT-C version A and Drawing Patterns Task B. According to Luria and Rosenblum (2012) and Smits-Engelsman and Van Galen (1997), these findings support the presumption that high load experienced due to high performance can lead to dis-automatization of writing processes. However, performance was not related to pressure-based smartpen features.

Regarding hypotheses 4a and b, results revealed that mean and standard deviation of velocity were higher for the low-ICL version of TMT-C. The latter conflicts the hypothesis and our finding that low variation of velocity was associated with higher performance within the task versions (hypotheses 3a and b). Possibly, smoothness of sketching velocity can represent different types of load. If the induced ICL remains the same (within versions), it may depend more on conscious effort how well a person performs. Smooth sketching might reflect both: strong effort, which is more associated with GCL than with ICL (Klepsch & Seufert, 2020) or high induced ICL. As expected, standard deviation of pressure was higher for the high-ICL version of the TMT-C, reflecting dis-automatization of the writing process (Luria & Rosenblum, 2012).

Results for hypothesis 5a indicated that smartpen measures could not predict children's subjective ICL ratings. Regarding hypotheses 5b, only mean velocity was identified as predictor for ECL. High mean velocity was related to low-ECL ratings. These results could partly be explained by missing metacognitive skills of children and the corresponding challenge of self-assessment (Chambers & Johnston, 2002) and by limited variance of ICL and ECL ratings. However, one can also question whether the two measures actually represent the same constructs, or maybe different ones (smartpen measures possibly rather effort), both related to performance.

Limitations and future research

This study adds to the limited literature on cognitive load measurement in middle childhood (Krieglstein et al., 2022), but it has several shortcomings regarding the generalizability of the findings.

The first limitation concerns the sample size. Particularly (multiple) regression analyses require large samples to reach sufficient levels of power. This leads to the fact that sample sizes of many behavioural regression studies, including the present one, are too small (Green, 1991; Maxwell, 2000). However, even if the results of this study can only point in certain directions, they nonetheless contribute a further puzzle piece to cognitive load measurement research. Particularly regarding the target group of children, there has been a lack of studies so far. Following Maxwell (2000) and Schmidt (1992), sufficient power and precision can often only be achieved through research syntheses across multiple studies. The open data of this study facilitate the inclusion of the current results in future meta-regressions. Moreover, findings and instruments of this study can serve as possible starting points for future study designs.

We expected that the children would be able to rate their cognitive load differentially for the dimensions under investigation. Based on our data, we cannot finally determine whether the multidimensional cognitive load measurement was successful. Although variations in the task demands affected the ratings on the intended load dimension, they also influenced the other dimension. Since ICL and ECL can hardly be varied independently from each other (Krieglstein et al., 2022), future research should verify the factor structure of the instrument with a larger sample. Moreover, to provoke more pronounced effects for smartpen-based and subjective cognitive load measures, future studies should operationalize load variations using task versions that are extremely different in terms of mental workload.

Regarding generalizability, relatively strong task-specific adaptations of the wording would be necessary to use the questionnaire in other contexts. Whether the psychometric quality would be affected by this needs to be examined in future research.

It also remains to be tested whether the results for the smartpen measures can be generalized to other sketching tasks. However, it is encouraging that the measures of velocity and pressure exhibited similar effects in our standardized sketching tasks as in the according prior research on handwriting.

The identification of reliable smartpen measures for cognitive load also contributes to using smartpens as helpful tool for people with special needs. Smartpens measures have been used for various diagnostic purposes: they were found to indicate developmental coordination disorders (Rosenblum & Livneh-Zirinski, 2008), dysgraphia (Rosenblum & Dror, 2016), autism spectrum disorder (Rosenblum et al., 2019), Parkinson's disease (Drotár et al., 2016) and cognitive impairment (Prange & Sonntag, 2022). Since cognitive (over)load is closely linked to cognitive impairment, future research should bring together findings on smartpen-based cognitive load indicators and cognitive deficits. Further, future studies on smartpen-based cognitive load measures should consider that writing process measures are also influenced by interindividual prerequisites, deficits and special needs.

Recent research on computer mouse interaction as a related fine motor task also performed by hand found a link between task-irrelevant, rather unconscious mouse actions and cognitive workload (Cha & Min, 2022). Task-irrelevant behaviour can also appear during writing and sketching tasks, for example, in forms of meaningless scribbling during cognitive information processing. The predictive power of smartpen measures during task-irrelevant writing behaviour for cognitive states should be investigated in future studies. Moreover, Cha and Min (2022) point out that mouse behaviour might not only be influenced by cognitive but also by affective states. In line with this, research on computer mouse interaction (Schaaff et al., 2012) and writing process measures (Schrader & Kalyuga, 2020) clearly implies that pressure parameters are related to emotions. Future studies should investigate the transferability of these results for the target group of children to identify affective moderator variables and combine information on cognitive and affective states for multifactorial learner state analyses.

CONCLUSION

Our research has shown that younger children can self-assess cognitive load and report it on a subjective rating instrument. Digital ink, specifically, velocity and pressure measures, were substantially related to performance and difficulty of the task, but barely with the individual cognitive load ratings. Regarding practical applications, smartpens allow for an efficient and non-intrusive, real-time cognitive load measurement and precise performance tracking (Sepp, 2019). Therefore, smartpens are a promising technology to support adaptive learning systems and research on learning analytics in children.

AUTHOR CONTRIBUTIONS

Kristin Altmeyer: Conceptualization; formal analysis; investigation; methodology; project administration; writing – original draft. **Michael Barz:** Conceptualization; formal analysis; investigation; methodology; software; visualization. **Luisa Lauer:** Conceptualization; investigation; methodology; writing – review and editing. **Markus Peschel:** Conceptualization; funding acquisition; resources. **Daniel Sonntag:** Conceptualization; funding acquisition; resources. **Roland Brünken:** Conceptualization; funding acquisition; resources. **Sarah Malone:** Conceptualization; formal analysis; investigation; methodology; project administration; supervision; writing – original draft.

ACKNOWLEDGEMENTS

This research was funded by the German Federal Ministry of Education and Research, BMBF (Project GeAR, Grant number: 01JD1811). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

None to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study will be openly available on OSF (https://osf.io/p3e6d/).

ORCID

Kristin Altmeyer https://orcid.org/0000-0003-1835-4432
Michael Barz https://orcid.org/0000-0001-6730-2466
Luisa Lauer https://orcid.org/0000-0002-0015-0821
Markus Peschel https://orcid.org/0000-0002-1334-2531
Daniel Sonntag https://orcid.org/0000-0002-8857-8709
Roland Brünken https://orcid.org/0000-0001-6038-8746
Sarah Malone https://orcid.org/0000-0001-8610-2611

REFERENCES

Anmarkrud, Ø., Andresen, A., & Bråten, I. (2019). Cognitive load and working memory in multimedia learning: Conceptual and measurement issues. *Educational Psychologist*, *54*(2), 61–83. https://doi.org/10.1080/00461520.2018.1554484

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*(5), 389–400. https://doi.org/10.1016/j.learninstruc.2006.09.001

Ayres, P., & Paas, F. (2012). Cognitive load theory: New directions and challenges. *Applied Cognitive Psychology*, *26*(6), 827–832. https://doi.org/10.1002/acp.2882

Badarna, M., Shimshoni, I., Luria, G., & Rosenblum, S. (2018). The importance of pen motion pattern groups for semi-automatic classification of handwriting into mental workload classes. *Cognitive Computation*, *10*(2), 215–227. https://doi.org/10.1007/s12559-017-9520-2

Barz, M., Altmeyer, K., Malone, S., Lauer, L., & Sonntag, D. (2020). *Digital pen features predict task difficulty and user performance of cognitive tests*. Paper presented at the proceedings of the 28th ACM conference on user modeling, adaptation and personalization, Genoa, Italy. https://doi.org/10.1145/3340631.3394839

Berninger, V. (1991). Overview of "bridging the gap between developmental, neuropsychological, and cognitive approaches to reading". *Learning and Individual Differences*, *3*, 163–179. https://doi.org/10.1016/1041-6080(91)90006-M

Cha, G. E., & Min, B. C. (2022). *Correlation between unconscious mouse actions and human cognitive workload*. In CHI conference on human factors in computing systems (pp. 1-7).

Chambers, C. T., & Craig, K. D. (1998). An intrusive impact of anchors in children's faces pain scales. *Pain*, *78*(1), 27–37. https://doi.org/10.1016/S0304-3959(98)00112-2

Chambers, C. T., & Johnston, C. (2002). Developmental differences in children's use of rating scales. *Journal of Pediatric Psychology*, *27*(1), 27–36. https://doi.org/10.1093/jpepsy/27.1.27

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement*. Springer.

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, *25*(2), 315–324. https://doi.org/10.1016/j.chb.2008.12.020

Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., & Faundez-Zanuy, M. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. *Artificial Intelligence in Medicine*, *67*, 39–46.

Eitel, A., Kühl, T., Scheiter, K., & Gerjets, P. (2014). Disfluency meets cognitive load in multimedia learning: Does harder-to-read mean better-to-understand? *Applied Cognitive Psychology*, *28*(4), 488–501. https://doi.org/10.1002/acp.3004

Eysink, T. H. S., de Jong, T., Berthold, K., Kolloffel, B., Opfermann, M., & Wouters, P. (2009). Learner performance in multimedia learning arrangements: An analysis across instructional approaches. *American Educational Research Journal*, *46*(4), 1107–1149. https://doi.org/10.3102/0002831209340235

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, *26*(3), 499–510.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, *23*(1), 1–19. https://doi.org/10.1007/s10648-010-9150-7

Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior*, *27*(1), 99–105. https://doi.org/10.1016/j.chb.2010.06.025

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, *8*, 1997. https://doi.org/10.3389/fpsyg.2017.01997

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, *48*(1), 45–77. https://doi.org/10.1007/s11251-020-09502-9

Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: A comparison of different objective measures. *Instructional Science*, *45*(4), 515–536. https://doi.org/10.1007/s11251-017-9413-5

Korbach, A., Ginns, P., Brünken, R., & Park, B. (2019). Should learners use their hands for learning? Results from an eye-tracking study. *Journal of Computer Assisted Learning*, *1-12*, 102–113. https://doi.org/10.1111/jcal.12396

Krieglstein, F., Beege, M., Rey, G. D., Ginns, P., Krell, M., & Schneider, S. (2022). A systematic meta-analysis of the reliability and validity of subjective cognitive load questionnaires in experimental multimedia learning research. *Educational Psycholy Review*, *34*, 2485–2541. https://doi.org/10.1007/s10648-022-09683-4

Leahy, W. (2018). Case studies in cognitive load measurement. In R. Z. Zheng (Ed.), *Cognitive load measurement and application: A theoretical framework for meaningful research and practice* (pp. 199–223). Routledge/Taylor & Francis Group.

Lee, H., & Mayer, R. E. (2015). Visual aids to learning in a second language: Adding redundant video to an audio lecture. *Applied Cognitive Psychology*, *29*(3), 445–454. https://doi.org/10.1002/acp.3123

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

Leppink, J., Paas, F., Van Gog, T., Van der Vleuten, C. P. M., & Van Merrienboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*(2014), 32–42. https://doi.org/10.1016/j.learninstruc.2013.12.001

Lin, T., Xie, T., Chen, Y., & Tang, N. (2013). Automatic cognitive load evaluation using writing features: An exploratory study. *International Journal of Industrial Ergonomics*, *43*(3), 210–217. https://doi.org/10.1007/s12559-015-9343-y

Luria, G., & Rosenblum, S. (2012). A computerized multidimensional measurement of mental workload via handwriting analysis. *Behavior Research Methods*, *44*(2), 575–586. https://doi.org/10.3758/s13428-011-0159-8

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, *5*(4), 434–458.

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning*, *8*(1), 19–46. https://doi.org/10.1007/s11409-013-9094-7

Oviatt, S., Hang, K., Zhou, J., Yu, K., & Chen, F. (2018). Dynamic handwriting signal features predict domain expertise. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(3), 1–21. https://doi.org/10.1145/3213309

Oviatt, S., Lin, J., & Sriramulu, A. (2021). I know what you know: What hand movements reveal about domain expertise. *ACM Transactions on Interactive Intelligent Systems*, *11*(1), 1–26. https://doi.org/10.1145/3423049

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Park, B., Korbach, A., & Brünken, R. (2015). Do learner characteristics moderate the seductive-details-effect? A cognitive-load-study using eye-tracking. *Journal of Educational Technology & Society*, *18*, 24–36.

Prange, A., & Sonntag, D. (2022). Modeling users' cognitive performance using digital pen features. *Frontiers in Artificial Intelligence*, *5*, 787179. https://doi.org/10.3389/frai.2022.787179

Ransdell, S., & Levy, C. M. (1999). Writing, reading, and speaking memory spans and the importance of resource flexibility. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing: Processing capacity and working memory in text production* (pp. 99–113). Amsterdam University Press.

Reitan, R. M. (1992). *Trail making test: Manual for administration and scoring*. Reitan Neuropsychology Laboratory.

Rheem, H., Verma, V., & Becker, D. V. (2018). Use of mouse-tracking method to measure cognitive load. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, pp. 1982–1986). SAGE Publications.

Rosenblum, S., Ben-Simhon, H. A., Meyer, S., & Gal, E. (2019). Predictors of handwriting performance among children with autism spectrum disorder. *Research in Autism Spectrum Disorders*, *60*, 16–24.

Rosenblum, S., & Dror, G. (2016). Identifying developmental dysgraphia characteristics utilizing handwriting classification methods. *IEEE Transactions on Human-Machine Systems*, *47*(2), 293-298. https://doi.org/10.1109/THMS.2016.2628799

Rosenblum, S., & Livneh-Zirinski, M. (2008). Handwriting process and product characteristics of children diagnosed with developmental coordination disorder. *Human Movement Science*, *27*(2), 200–214.

Rosenblum, S., & Luria, G. (2016). Applying a handwriting measurement model for capturing cognitive load implications through complex figure drawing. *Cognitive Computation*, *8*(1), 69–77. https://doi.org/10.1007/s12559-015-9343-y

Ruiz, N., Taib, R., Shi, Y., Choi, E., & Chen, F. (2007). *Using pen input features as indices of cognitive load*. In Proceedings of the 9th international conference on multimodal interfaces (pp. 315–318). https://doi.org/10.1145/1322192.1322246

Schaaff, K., Degen, R., Adler, N., & Adam, M. T. (2012). Measuring affect using a standard mouse device. *Biomedical Engineering*, *57*, 761–764.

Schmeck, A., Opfermann, M., van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*(1), 93–114. https://doi.org/10.1007/s11251-014-9328-3

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*(10), 1173–1181.

Schrader, C., & Kalyuga, S. (2020). Linking students' emotions to engagement and writing performance when learning Japanese letters with a pen-based tablet: An investigation based on individual pen pressure parameters. *International Journal of Human-Computer Studies*, *135*, 102374.

Schroeder, N. L., & Cenkci, A. T. (2018). Spatial contiguity and spatial Split-attention effects in multimedia learning environments: A meta-analysis. *Educational Psychology Review*, *30*(3), 679–701. https://doi.org/10.1007/s10648-018-9435-9

Sepp, S. (2019). *Meaningful hand gestures for learning with touch-based I.C.T.* Doctor of Philosophy thesis, School of Education, University of Wollongong. https://ro.uow.edu.au/theses1/903

Skulmowski, A., & Rey, G. D. (2020). Subjective cognitive load surveys lead to divergent results for interactive learning media. *Human Behavior and Emerging Technologies*, *2*(2), 149–157. https://doi.org/10.1002/hbe2.184

Smits-Engelsman, B. C. M., & Van Galen, G. P. (1997). Dysgraphia in children: Lasting psychomotor deficiency or transient developmental delay? *Journal of Experimental Child Psychology*, *67*, 164–184. https://doi.org/10.1006/jecp.1997.2400

Snijders, P., Tellegen, P. J., & Laros, J. A. (2005). *SON-R 5½-17. Snijders-Oomen-Non-Verbal intelligence test* (3rd ed.). Hogrefe.

Stebner, F., Kühl, T., Höffler, T. N., Wirth, J., & Ayres, P. (2017). The role of process information in narrations while learning with animations and static pictures. *Computers & Education*, *104*, 34–48. https://doi.org/10.1016/j.compedu.2016.11.001

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

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Intrinsic and extraneous cognitive load. In *Cognitive load theory. Explorations in the learning sciences, instructional systems and performance technologies* (Vol. 1). Springer. https://doi.org/10.1007/978-1-4419-8126-4_5

Sweller, J., van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tang, M., Ginns, P., & Jacobson, M. J. (2019). Tracing enhances recall and transfer of knowledge of the water cycle. *Educational Psychology Review*, *31*, 439–455. https://doi.org/10.1007/s10648-019-09466-4

Van Gemmert, A. W. A., & Van Galen, G. P. (1998). Auditory stress effects on preparation and execution of graphical aiming: A test of the neuromotor noise concept. *Acta Psychologica*, *98*, 81–101. https://doi.org/10.1016/S0001-6918(97)00049-8

Willems, D., & Niels, R. (2008). *Definitions for features used in online pen gesture recognition*. NICI, Radboud University Nijmegen.

Witte, T. E., Haase, H., & Schwarz, J. (2021). *Measuring cognitive load for adaptive instructional systems by using a pressure sensitive computer mouse*. In International conference on human-computer interaction (pp. 209–218). Springer, Cham.

Wu, Z., Lin, T., & Tang, N. (2016). Explore the use of handwriting information and machine learning techniques in evaluating mental workload. *International Journal of Technology and Human Interaction (IJTHI)*, *12*(3), 18–32. https://doi.org/10.4018/IJTHI.2016070102

Yu, K., Epps, J., & Chen, F. (2011). *Cognitive load evaluation of handwriting using stroke-level features*. In Proceedings of the 16th international conference on intelligent user interfaces (pp. 423–426). https://doi.org/10.1145/1943403.1943481

Yung, H. I., & Paas, F. (2015). Effects of computer-based visual representation on mathematics learning and cognitive load. *Journal of Educational Technology & Society*, *18*(4), 70–77.

Zu, T., Hutson, J., Loschky, L. C., & Rebello, N. S. (2020). Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *Journal of Educational Psychology*, *112*(7), 1338–1352. https://doi.org/10.1037/edu0000441

How to cite this article: Altmeyer, K., Barz, M., Lauer, L., Peschel, M., Sonntag, D., Brünken, R., & Malone, S. (2022). Digital ink and differentiated subjective ratings for cognitive load measurement in middle childhood. *British Journal of Educational Psychology*, *00*, 1–18. https://doi.org/10.1111/bjep.12595