Tracing the learning experience in a digital learning environment: insights from pen pressure and facial expressions in primary school children

Donna Bryce, Jana Spear, Cara-Sophie Enste, Robert Grassinger, Markus Dresel

Angaben zur Veröffentlichung / Publication details:

Bryce, Donna, Jana Spear, Cara-Sophie Enste, Robert Grassinger, and Markus Dresel. 2026. "Tracing the learning experience in a digital learning environment: insights from pen pressure and facial expressions in primary school children." *Metacognition and Learning* 21 (1): 32. https://doi.org/10.1007/s11409-026-09482-0.

Metacognition and Learning (2026) 21:32
https://doi.org/10.1007/s11409-026-09482-0

Tracing the learning experience in a digital learning environment: Insights from pen pressure and facial expressions in primary school children

Donna Bryce[1] · **Jana Spear**[1] · **Cara-Sophie Enste**[2] · **Robert Grassinger**[2] · **Markus Dresel**[1]

Received: 29 October 2025 / Accepted: 11 June 2026

Abstract
Learning independently involves not only cognitive but also emotional, motivational, and metacognitive challenges, particularly for children still developing their self-regulated learning skills. Consequently, children's learning experiences can fluctuate rapidly within one learning activity, and assessing these experiences without disrupting learning remains an important research aim. In the present study, the feasibility of unobtrusively collecting two forms of trace data—pen pressure and facial expressions—from young learners in authentic classroom contexts was investigated, and the validity of the resulting indicators was tested. Data was analysed from two classroom studies with over 580 third- and fourth-grade children in Germany, who learned mathematics content in a tablet-based digital learning environment using digital pens and integrated webcams. Pen pressure was recorded in three phases of the learning task, facial expressions following feedback were coded for valence, and linear mixed-effects models tested how these indicators varied with item accuracy and difficulty. Mean and variance in pen pressure increased with item difficulty and after positive feedback, suggesting pen pressure indicators reflect both cognitive load and arousal. More varied pen pressure on difficult items that were answered incorrectly may reflect the metacognitive experience of uncertainty. Children's facial expressions were rated more positively after positive feedback than error feedback and when they self-reported higher valence, while nuanced effects of item difficulty pointed to attributional processes shaping children's emotional reactions. Overall, the findings demonstrate that pen pressure and facial expressions can be collected unobtrusively in real classroom contexts and provide valuable insights into young learners' experiences.

Keywords Self-regulated learning · Trace data · Pen pressure · Facial expression · Children · Mathematics

Extended author information available on the last page of the article

Published online: 02 July 2026

Introduction

Learning new topics and acquiring new strategies for solving complex tasks is not only a cognitive challenge, but also an inherently emotional, motivational, and metacognitive challenge for learners of all ages. For children, who are still transitioning from other- to self-regulating their learning (Cole et al., 2019), self-study can be especially challenging and dynamic. While processing learning content, learners can experience emotions like pride and boredom (e.g., Pekrun et al., 2016), be motivationally interested or not value the topic (e.g., Renninger et al., 2014; Wigfield & Eccles, 2000), and be affected by metacognitive experiences such as confidence and uncertainty (Efklides, 2006). As they begin to apply a new strategy, for example in some self-test items, learners may evaluate the difficulty of an item, invest variable degrees of effort, and feel varying levels of certainty and frustration about their answers. When they receive feedback on their newly acquired skills, learners may experience joy, relief, disappointment or shame, and the feedback may concur with their prior expectations of success or not (e.g., Horvers et al., 2025). As is clear from these examples, a myriad of experiences can arise and fluctuate rapidly during different phases of learning. A pressing challenge for researchers committed to capturing the wide spectrum of learning phenomena—including cognitive, emotional, motivational, and metacognitive processes—is the need to collect high-frequency data on these processes as unobtrusively as possible. The current study aims to contribute to addressing this challenge.

Assessing self-regulated learning processes dynamically

All dominant models of self-regulated learning (SRL) acknowledge the dynamic and cyclical nature of the processes and behaviours that contribute to learning success (Panadero, 2017). For instance, Boekaerts' (2011) Dual-Processing Model emphasises that whether a learner finds themselves on a growth or a well-being pathway is influenced by their appraisals of the learning task and their associated learning intentions; importantly, these appraisals are generated repeatedly throughout a learning task. The frameworks by Pintrich (2000) and Zimmerman (2000) arrange processes relevant to self-regulated learning in three or four temporal phases, but emphasise in their descriptions that, "there is no strong assumption of a simple linear, static process with separable noninteracting components" (Pintrich, 2000, p. 456). Despite this, much research conducted into self-regulated learning processes has employed rather static methods, with a strong reliance on self-report regarding how one typically behaves, or how one behaved in a specific learning task (Rovers et al., 2019; Winne, 2010). Such reports can be limited in the temporal dimension, with memory biases being unavoidable when a time delay is present. While some designs improve on this by collecting reports frequently and proximally to the learning process, for instance ambulatory assessment or experience sampling, it is generally agreed that self-reports are limited in their ability to evaluate dynamically changing experiences during learning and can be intrusive if implemented with high frequency (Rintala et al., 2023). Further, there are well documented concerns about children's ability to provide reliable self-reports (Conjin et al., 2020).

Many of the aforementioned limitations of self-report also apply to parent- or teacher-reports of children's self-regulation or self-regulated learning abilities such as the Children's Behavior Questionnaire (Rothbart et al., 2001) and the CHILD (Whitebread et al., 2009). In addition, their validity has been questioned (Karlen et al., 2024) and external

reporters are naturally restricted in their ability to report on the inner emotional and motivational experiences of young learners. While progress has been made in the development and validation of observational tools that aim to unobtrusively capture the naturally occurring self-regulated learning abilities of young learners, e.g., the Regulation-Related Skills Measure (McCoy et al., 2022; Eberhart et al., 2024), these too are primarily based on observable behaviours and cannot access learners' inner experiences. Furthermore, they remain time- and resource-intensive to apply and may result in observer effects which threaten validity. As such, there seems to be a need for more objective, scalable, and minimally intrusive assessment methods in the field of self-regulated learning.

The collection of trace or log data is a promising option for gaining more insight into learners' dynamic inner experiences during learning tasks (Kovanovic et al., 2023). These types of data can include log data regarding how a learner navigates a learning environment, time on screen, response times, mouse-tracking trajectories, pressure applied with finger or digital pens, as well as video data collected from webcams. The analysis of trace data offers substantial advantages, including the capacity to capture fine-grained, unobtrusive records of learning processes. However, two methodological challenges warrant consideration. First, the high granularity of trace data requires careful theoretical grounding to link observed behavioural indicators to specific latent self-regulation constructs (Seufert, 2026; Winne, 2010). Second, validation of trace data interpretations benefits from triangulation with other measures. Concurrent think-aloud protocols have been successfully combined with log data in upper primary school children (Paans et al., 2019; Vandevelde et al., 2015). However, such protocols require substantial resources and may not be feasible in all research or educational contexts. As such, complementary validation approaches are valuable, such as examining how trace data indicators systematically vary with item characteristics or context variables. This is the approach taken in the current study.

Trace data are typically collected within a digital learning environment. One notable example in the self-regulated learning field is MetaTutor (Azevedo et al., 2022) which has the dual aims of promoting and assessing self-regulated learning processes multimodally. MetaTutor was developed for late adolescent and adult learners and has primarily been studied under laboratory conditions. There remains a paucity of programs designed to collect trace data from younger learners (namely, primary school-aged learners) in ecologically valid contexts. Indeed, in a recent review of trace data indicators of self-regulated learning, only six studies included participants from primary, middle or high school (Boulahmel et al., 2025). With increased use of digital tools in education, more opportunities arise for researchers and practitioners to access and make use of trace data using technology freely available in regular classrooms. As primary school classrooms are increasingly equipped with tablet computers or easy access to these (OECD, 2023), in the current study we designed a digital learning environment for tablet computers. These devices offer easy data collection of two types of trace data that have been previously linked to cognitive, motivational, and emotional processes, namely pen pressure (collected via a digital pen) and facial expression ratings (based on videos collected via integrated cameras). The overarching aim of the current study is to test the feasibility of collecting trace data from primary school-aged learners in real-life learning contexts and evaluate the validity of the resulting variables.

Pen pressure

The majority of prior studies investigating pen pressure in learning contexts have been conducted with small samples, adult participants, and under controlled laboratory conditions. As such, they are only of limited relevance to our current aim to investigate primary school students in authentic settings. Further, the studies are quite heterogenous in terms of equipment used (digital pens on real paper or on tablets), type of task (academic or game-based), and the actions during which pen pressure is measured (writing, drawing, tapping on buttons). Nevertheless, findings do seem to indicate that more pressure and more variable pressure is associated with more difficult conditions and/or higher cognitive load. For example, Li & colleagues (2024) measured pen pressure (including mean and standard deviation of pressure) while adult participants engaged with two learning tasks—writing sentences in a second language, and taking notes and highlighting text during text comprehension. The data indicated that pen pressure alone was not a particularly strong predictor of participant-reported item difficulty, but its inclusion did improve the predictive models (i.e., it is one of many meaningful multimodal data sources, along with stroke-based, path-based, time-based, and eye-tracking features). In a study with 7- to 12-year-old children, Altmeyer and colleagues (2023) measured pen pressure while children completed two sketching tasks that varied in cognitive load. While pen pressure was not associated with child-reported cognitive load, children applied more variable pen pressure in a trail making task with higher intrinsic cognitive load than with low intrinsic cognitive load. Aside from predicting cognitive load or task difficulty, some researchers have tried to predict high school students' domain expertise based only on pen measures collected while they worked on complex maths problems (Oviatt et al., 2018). Results showed that stroke distance and mean pressure were the greatest negative predictors of maths expertise; that is, students who could more easily solve the problems used shorter pen strokes and applied less pressure while working than their peers who struggled. Despite these studies predicting slightly different outcome variables, in sum, they suggest that both the average applied pen pressure as well as the variability thereof are positively related to item complexity or cognitive load.

Lastly, pen pressure has also been associated with emotion and motivation variables, with higher mean pressure correlating positively with self-reported frustration, negatively with enjoyment and negatively with task engagement in adults learning to write Japanese letters (Schrader & Kalyuga, 2020). Given the paucity of studies examining pen pressure in primary school children, it is not known to what extent this indicator will reflect the same or different processes as in adult participants. Further, whether the associations evidenced in data collected under rather controlled conditions will generalise to data collected in a more ecologically valid classroom setting remains to be seen.

Facial expressions

The facial expressions learners make while learning can be observed and rated for emotion. There are two main approaches to this—human rating based on predefined coding schemes (conducted in situ or with videos, e.g., Somerville & Whitebread, 2019; Horvers et al., 2025; Merrick & Fyfe, 2023) and automated systems that analyse video data (e.g., FaceReader, Noldus Information Technology, 2022). In both approaches, discrete emotions (e.g., joy, anger, boredom, surprise), valence (positive to negative), and/or arousal (low to high) can be

rated, depending on the research aims. While automatic systems are undoubtedly attractive because of their time efficiency, they have not been widely validated in children (Cross et al., 2023) and remain inaccessible to many researchers because of high costs and required expertise. Consequently, in this study human coding was the chosen method for rating facial expressions in young children observed in real classroom contexts.

Ratings of facial expressions have primarily been validated by calculating their concurrence with participants' self-reported emotions. Generally speaking, in learning contexts children's self-reported emotions tend to be rather positive and their facial expressions tend to be rated more neutrally (Horvers et al., 2025) and in adults it has been established that fewer emotions are observable than are reported (Reisenzein et al., 2014). In a study with 10- to 12-year-olds significant but small correlations emerged between researcher-observed negative emotions and self-reported valence, and no correlations emerged between researcher-observed positive emotions and self-reported valence (Horvers et al., 2025). The masking of emotional expressions, i.e., when an individual quickly changes their facial expression to hide their experienced emotion, may pose a threat to the validity of facial expression ratings. This may be particularly pronounced in interactions, where facial expressions serve a social and communicative function. Indeed, in a naturalistic observational study in primary school classrooms, Somerville and Whitebread (2019) established that the majority of emotionally challenging situations involved an interaction with a peer. It is conceivable that negative emotion expressions are masked more in social interactions or when learners feel like they are being observed, in comparison to self-study contexts when facial expressions are captured unobtrusively via video recordings. Additionally, we speculated that social display rules may not be as established or pronounced in younger learners. For these reasons, in the current study we were optimistic that emotional reactions would be observable when children received feedback on their performance.

Two prior studies have specifically examined primary school students' facial expressions following feedback on maths problems (Horvers et al., 2025; Merrick & Fyfe, 2023). Horvers and colleagues coded for discrete emotions and observed that more negative than positive emotions were observed following negative feedback, but a similar frequency of negative and positive emotions were observed following positive feedback. Merrick and Fyfe rated children's emotional valence both while they worked on maths problems and after they received feedback; they based their ratings on facial expressions, tone of voice as well as verbal statements. Similar to Horvers et al. (2025), they observed more negative valence following negative feedback than positive feedback, and also that more positive valence was observed in feedback than problem-solving phases, but that negative valence was similarly prevalent across the phases. The latter finding is presumably rather task specific and related to the difficulty of items. Indeed, a study on the impact of game elements on facially expressed emotions and subjective effort found that associations between facial expressions and task conditions are not always consistent suggesting that what emotions are observed and what conclusions can be drawn is very context-dependent (Greipl et al., 2021). For this reason, in the current study we analyse two independent datasets and highlight commonalities in results. In terms of effort and task engagement, studies from the field of game-based learning indicate that facial expressions reflecting a broad spectrum of positive and negative valence and high and low arousal are associated with subjective effort (Greipl et al., 2021) and that more engaging game-based tasks are associated with higher proportions of observed happiness and sadness (Ninaus et al., 2019). What remains unknown, is

whether facial expressions in response to feedback are influenced by the effort required to solve specific problems, i.e., by item difficulty.

The current study

The present study addresses the need for objective, scalable, and unobtrusive methods to capture young learners' dynamic experiences during self-study. Using a tablet-based digital learning environment, two forms of trace data were collected—pen pressure via digital pens and facial expression ratings based on video recordings—from primary school-aged children in real classroom settings. While previous research has linked pen pressure and facial expressions to cognitive load, task difficulty, motivation, emotion, and effort, much of this work has been conducted with adults and under controlled laboratory conditions, leaving open questions about their validity and usefulness in authentic classroom contexts with younger learners. By examining how three indicators derived from pen pressure measurement and facial expression ratings are affected by response accuracy and item difficulty, this study evaluates the potential of these trace data sources to provide meaningful indicators of children's cognitive, motivational, and emotional processes. In our approach, variables derived from trace data are treated as dependent variables, and features of the learning environment are treated as predictors in order to estimate their unique contributions efficiently while avoiding multiple testing and redundancy across separate models.

In a first research question, we investigated what predicts the mean pen pressure applied during the learning environment. Hereby we considered the predictors accuracy of answer, item difficulty, and the phase within learning environment. Based on prior findings regarding cognitive load and pen pressure (namely, Li et al., 2024 and Oviatt et al., 2018), we expected that when participants were typing and entering their answers, they would apply more pressure for difficult than easy items, and for items that they answered incorrectly compared to correctly. Despite a smaller relevant evidence base, one might expect higher mean pen pressure following negative than positive feedback, based on Schrader and Kalyuga's (2020) findings regarding frustration and enjoyment.

We also tested what predicts the variance of pen pressure applied during the learning environment (research question 2). Hereby we considered the predictors accuracy of answer, item difficulty, and the phase within learning environment. Based on the prior findings of Altmeyer and colleagues (2023), we expected more variable pressure when participants were typing and entering their answers for more difficult than easier items, and in incorrectly than correctly responded items. Since Schrader and Kalyuga (2020) did not investigate the variance of pen pressure as a dependent variable, no firm hypotheses were made regarding variance of pen pressure applied after receiving feedback.

Lastly, in a third research question it was investigated what predicts facial expression valence ratings after receiving feedback. Hereby we considered the predictors accuracy of answer and item difficulty. Based on two relevant studies (namely, Horvers et al., 2025 and Merrick & Fyfe, 2023), we expected more positive valence to be expressed for correctly responded than incorrectly responded items; no strong hypotheses were made regarding the effect of item difficulty. Additionally, in an initial validation step we investigated the concurrence between children's self-reported valence and arousal on the one hand, and the facial expression valence ratings made by external observers on the other hand. We expected to

observe significant but low concurrence with participants' self-reported valence, and no associations with participants' self-reported arousal.

Methods

The datasets analysed in the current manuscript are subsets of data collected as part of two larger studies within primary schools in Germany. Participation was voluntary and parents/guardians provided written consent. Additionally, parents could opt in or out of video data being recorded. The study's procedures fully complied with relevant laws and institutional guidelines, and ethical approval was granted by the ethics committee of the University of Education Weingarten (on 1st August 2023, reference number: PHW20230801) and the regional government (on 4th August 2023, reference number: 40.1-5038-2/6). Here only data from participants who agreed to provide video data and whose data included at least one correctly and one incorrectly responded trial are analysed. Analysis of a subset of the video data from Study 1 is reported in Pickal et al. (2026). The Study 2 dataset contains data from an experimental and control group, in which children's beliefs about errors were manipulated. Since none of the variables of interest that are analysed here were affected by the experimental manipulation, data from both groups are included.

Participants

Data analysed from Study 1 were provided by $N=243$ third grade children attending 21 classes from 6 primary schools in Germany. The children were on average 9.28 years old ($SD=0.37$), 48% were male, 49% were female and 2% were non-binary. The mean of their last grade in Maths was 2.25, which is classified as "Good to Satisfactory" in the German system which grades from 1=*very good* to 6=*fail*. 17% of the sample were not born in Germany, and 45% had at least one parent not born in Germany. Data analysed from Study 2 were provided by $N=341$ third and fourth grade children attending 24 classes from 7 primary schools in Germany. The children were on average 9.98 years old ($SD=0.62$), 49% were male, 49% were female and 1% were non-binary. The mean of their last grade in Maths was 2.26, which is classified as "Good to Satisfactory" in the German system. 22% of the sample were not born in Germany, and 53% had at least one parent not born in Germany. With the percentage of adults in Germany with an immigration background currently around 30% (Statistisches Bundesamt, 2025) this sample can be considered representative of the broader population.

Materials and apparatus

Data were collected from children within a digital learning environment, programmed using PsychoPy (Version: 2024.2.4; Peirce et al., 2019) and hosted online via www.pavlovia.org (Bridges et al., 2020). Children processed the digital learning environment using tablet computers equipped with webcams (Microsoft Surface Pro 4), headphones (TIMIO children's headphones) and a digital pen (Microsoft Surface Pen). The tablet computers were positioned upright in front of the children, using the integrated stand in the back of the device. Accordingly, the videos recorded from the built-in webcams generally aligned

with the children's faces, except when the children shifted away from their typical posture. The pressure applied via the digital pen was recorded using the API PointerEvent; a demonstration of how this can be implemented in Pavlovia is freely accessible and shared online (Papenmeier, 2024). Pressure was measured as a normalized value (range: 0–1) via the PointerEvents API, where values represent the proportion of maximum detectable pressure for each device (unitless).

For both Study 1 and Study 2, the content of the digital learning environment was mathematics; the topics and quiz items selected varied slightly to ensure that they were curriculum-relevant for the children at the time of data collection. Consequently, in Study 2, different learning materials were presented to the third- and fourth grade students. A full list of chapter names and quiz items is provided in Supplementary Materials. In order to validate the coding scheme used to rate valence of facial expressions, the valence and arousal scales of the Self-Assessment Manikin (SAM; Bradley & Lang, 1994) were collected from children within the digital learning environment. The ratings children provided on the SAM were reverse coded so that for Valence 1 = *negative* and 9 = *positive valence*, and for Arousal 1 = *low* and 9 = *high arousal*.

Procedure

The data were collected in schools, in a group data collection session during a lesson in a typical school day. Children worked on the content of the digital learning environment in a self-paced manner under quiet study conditions (for 30 min in Study 1 and for 40 min in Study 2). One trained researcher conducted the session, occasionally accompanied by the class teacher, who was instructed not to assist the children in answering the quiz items. Children were instructed to use the digital pen when interacting with the tablet. Children exited the digital learning environment after the fixed amount of time, resulting in a different number of completed chapters and quiz items per participant. Students were informed that completing all four chapters of the digital learning environment was not expected or required; rather, the emphasis was placed on engaging in effective learning.

Early in the digital learning environment, children completed five "practice" slide actions with their digital pen on the tablet surface (see A in Fig. 1); these served as calibration trials from which mean and standard deviation (*SD*) of pen pressure was calculated and used as a control variable in models with the dependent variables mean pen pressure or *SD* pen pressure, respectively. It was deemed important to assess pen pressure outside of the learning task and control for the relevant variables in our analyses because there could be individual differences in children's experience using digital pens on tablet computers as well as in their general motor control (see Potamianou & Bryce, 2026, for a similar approach in a mouse-tracking study with children). Afterwards, they worked through up to four chapters, each targeting a different curriculum-relevant topic for their year group and the timepoint within the school year. Each chapter had the same structure: first, between six and ten pages of content were presented to the children regarding strategies for solving specific types of maths problems (see "Chapter content" in Fig. 1). This information was presented audio-visually and the children could re-play audio explanations and navigate forwards and backwards through these pages as they wished. After each chapter content, five quiz items with feedback were presented one after the other. Children solved the items (using the note paper if desired), typed their answers by tapping on a number pad presented digitally on the screen

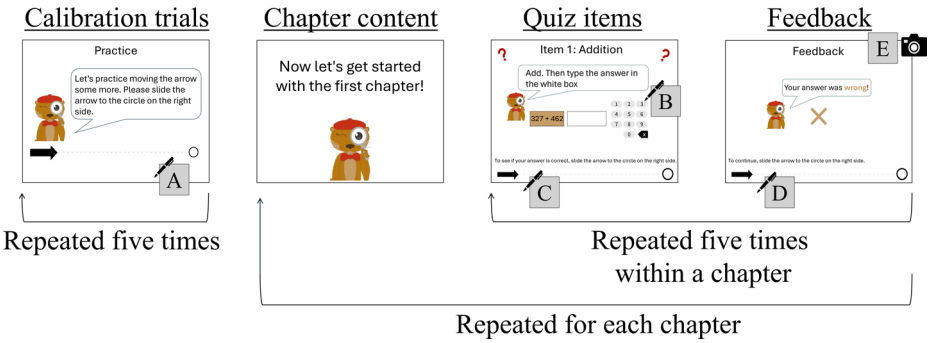

Fig. 1 Schematic of the procedure and analysed variables. Note. Mean calibration pressure and *SD* calibration pressure were calculated for each participant from five calibration trials (see **A**). Phase 1 Typing Answer values were calculated from the pressure measured when the participant tapped on each button of the digital number pad (see **B**). Phase 2 Entering Answer values were calculated from the pressure measured when the participant moved the arrow from left to right on a Quiz Item screen (see **C**). Phase 3 After Feedback values were calculated from the pressure measured when the participant moved the arrow from left to right on a Feedback screen (see **D**). Video clips of facial expressions were collected during each Feedback screen (see **E**) and coded for emotional valence.

(see B in Fig. 1), and entered their answer by sliding an arrow presented at the bottom of the screen from left to right (see C in Fig. 1). Immediately after entering the answer, the children received feedback on their answer to that quiz item (see "Feedback" in Fig. 1). During this time, short videos were recorded using the tablet-integrated webcam (see E in Fig. 1). To proceed from the feedback presentation, children had to again slide an arrow presented at the bottom of the screen from left to right (see D in Fig. 1). Children's self-reported valence and arousal was collected using the SAM immediately after item feedback in chapter 1 only (Study 1) or in all chapters (Study 2).

After completing the five quiz items at the end of a chapter, summary feedback containing information about whether each item was answered correctly or incorrectly was presented to the children. Next, they could decide whether to repeat the chapter a second time or proceed to the next chapter. Only data from the first attempts at each chapter are included in the current analyses, since different physiological reactions may occur when encountering a quiz item for a second time.

Data preparation and analyses

Single trial exclusion

All trials with video data, those in which the participant provided an answer to the quiz item, and those that were the first attempt at each quiz item were eligible for analyses. This amounted to 3263 trials in Study 1 and 4581 trials in Study 2. Some single trials were excluded from each dataset for specific technical reasons. Trials with missing pressure data in one of the three phases, probably because the child used their finger instead of the digital pen, were excluded (in Study 1 this was 3% of all trials; in Study 2 this was 8% of all trials). This resulted in 3174 trials from 243 participants in Study 1 and 4209 trials from 341 participants in Study 2 to address research questions 1 (What predicts mean pen pressure?) and 2 (What predicts the variance of pen pressure?). For the analysis to address research ques-

tion 3 (What predicts facial expression?), only trials with codable videos could be included. This further reduced the datasets to 2055 trials from 200 participants in Study 1 and 2991 trials from 317 participants in Study 2. Videos were deemed not codable if: the face was not clearly visible from the forehead to bottom lip, a hand covered parts of the mouth or the child's fingers were in their mouth, lighting was so bad that the face was not clearly visible, a hand covered the eyes, or an unnatural situation occurred such as the teacher talking to the student.

Calibration trials

As described in Procedure, each child completed five calibration trials at the start of the digital learning environment. Single calibration trials with missing pressure data (9 trials from Study 1; 38 trials from Study 2; most likely due to a finger being used instead of the digital pen) were excluded from further analyses and calibration variables were derived from the remaining calibration trials. Pen pressure information was recorded every time the screen was refreshed while the digital pen interacted with the tablet screen. As such, to calculate the calibration mean pen pressure for each participant, first the mean pressure applied within each calibration trial was calculated, and then the mean of these (up to) five values was calculated. To calculate the calibration variance pen pressure for each participant, first the *SD* of pressure values within each calibration trial was calculated, and then the mean of these (up to) five values was calculated. One of these variables was entered as a control variable for models with the relevant pen pressure variable (mean or variance) as dependent variable.

Pen pressure variables

For each phase (see B, C, and D in Fig. 1), both the mean pen pressure and the variance of pen pressure was calculated, similarly as described above. That is, in the Typing Answer phase, the pressure with which a number button was tapped was recorded. The mean and *SD* of these values were calculated for Phase 1. In the Entering Answer phase and the After Feedback phase, the pressure with which the arrow was pressed as the participant slid it from left to right was recorded. The mean and *SD* of these values were calculated for Phase 2 and Phase 3, respectively.

Item difficulty

Item difficulty was operationalized using the proportion-correct method in Classical Test Theory (Lord & Novick, 2008), where lower accuracy within the sample reflects higher difficulty. For the 20 items included in Study 1, the item difficulty ranged from 0.07 to 0.65; for the 40 items included in Study 2 (20 items for third grade, 20 items for fourth grade), the item difficulty ranged from 0.14 to 0.87.

Facial expression valence rating

The valence of a child's facial expression as observed in the video clips was rated on a 5-point rating scale from *very negative* (1) to *very positive* (5). This rating was based on a coding scheme adapted from those by Cole et al. (1994), Somerville & Whitebread (2019),

and Merrick & Fyfe (2023). The full coding scheme can be found in Supplementary Materials. Two trained research assistants coded a subset of the same videos (52 in total) and reached very good interrater agreement (ICC = 0.82), before continuing to provide ratings for approximately half of the video clips each. The research assistants were blind to which item the student in the video had attempted and whether they had responded correctly or incorrectly.

Approach to data analysis

Data were analysed using R (R Core Team, 2025). To address the research questions, single trials were analysed in three linear mixed effects models (LMMs; using the package *lme4*, Bates et al., 2014), with the dependent variables mean pen pressure, variance pen pressure and valence rating, respectively. For the first two models, the associated control variable (calibration mean pen pressure, and calibration variance pen pressure) was entered as a fixed effect. Additional fixed effects for these two models were the Accuracy of the Answer (correct, incorrect), Item Difficulty, and Phase within the digital learning environment (1: Typing Answer, 2: Entering Answer, 3: After Feedback). For the model predicting valence rating from the video data, fixed effects included Accuracy of the Answer (correct, incorrect) and Item Difficulty. To aid interpretation, the Item Difficulty fixed effect was centred before being entered into the LMMs. For all models, a random intercept for Participant was included to account for individual differences in the baseline dependent variables, acknowledging that repeated trials from the same participant are not independent. For models with pen pressure variables as dependent variables, the hierarchical structure of the data was modelled by also including random intercepts for the interaction between participant and trial (as pen pressure data from three phases within one trial were included as fixed effects). For each dependent variable, the best fitting LMM was established via the model selection procedure described by Barr and colleagues (2013). The full models include all predictors as simple effects and their higher order interactions. In an iterative fashion, the impact of removing one term is tested using likelihood ratio tests (using the package *lmerTest*, Kuznetsova et al., 2017) and the best fitting LMMs are presented in tables.

Results

Descriptive statistics

Table 1 contains the group means, calculated from the participant means, for key variables collected within the digital learning environment for each study. It appears that children in Study 2 progressed through more chapters, performed overall less accurately, took longer to answer the questions and spent longer reading their feedback than did children in Study 1. They also reported more neutral (than positive) valence and higher arousal after reading their feedback, applied more pen pressure in Phase 1 and 3, and more variable pressure in Phase 1 than did children in Study 1.

Table 1 Summary of participant means of each relevant variable

	Study 1				Study 2			
	N	Mean	*SD*	Min–Max	*N*	Mean	*SD*	Min–Max
Calibration mean pressure	243	0.25	0.09	0.06–0.60	341	0.27	0.09	0.05–0.60
Calibration variance pressure	243	0.10	0.03	0.03–0.26	341	0.10	0.03	0.03–0.20
Behaviour in learning task								
Number of chapters completed	243	2.97	0.87	1.00–4.00	341	3.54	0.81	1.00–4.00
Overall response accuracy	243	0.72	0.21	0.00–1.00	341	0.49	0.27	0.00–1.00
Calculation time (s)	243	52.83	21.85	14.40–173.07	341	82.01	40.67	14.88–364.23
Feedback reading time (s)	243	3.88	2.16	1.74–24.73	341	4.79	4.08	1.75–41.15
Phase 1: Typing answer								
Mean pressure	243	0.19	0.06	0.06–0.40	341	0.21	0.07	0.06–0.44
Variance pressure	243	0.06	0.01	0.02–0.12	341	0.06	0.02	0.01–0.14
Phase 2: Entering answer								
Mean swipe pressure	243	0.29	0.09	0.11–0.68	341	0.30	0.10	0.06–0.56
Variance swipe pressure	243	0.11	0.03	0.04–0.24	341	0.11	0.03	0.03–0.25
Phase 3: After feedback								
Mean swipe pressure	243	0.29	0.09	0.11–0.67	341	0.31	0.10	0.08–0.60
Variance swipe pressure	243	0.11	0.03	0.05–0.26	341	0.11	0.03	0.04–0.20
Facial expression coding	200	2.95	0.35	2.00–4.88	317	2.94	0.45	1.50–5.00
Self-reported valence	243	7.80	1.47	1.80–9.00	341	6.27	2.04	1.00–9.00
Self-reported arousal	243	2.75	1.94	1.00–9.00	341	4.20	2.23	1.00–9.00

Table 2 Estimates for fixed effects in the LMMs with facial expression valence rating as dependent variable for Study 1 and Study 2

Predictor	Study 1			Study 2		
	Estimate	95% CI	*p*	Estimate	95% CI	*p*
(Intercept)	3.05	2.98–3.11	<0.001	2.94	2.90–3.00	<0.001
Self-reported valence[a]	0.09	0.06–0.12	<0.001	0.08	0.07–0.09	<0.001
Self-reported arousal[a]	−0.002	-0.03–0.02	0.872	0.001	−0.01–0.01	0.919

p-values are based on the Wald statistic

LMM Linear Mixed-Effects model, *CI* Confidence interval

[a]variable was mean centred before entering model

Validation of facial expression valence rating

To investigate the validity of our rating of facial expression valence, it was tested whether the children's self-reported valence was a significant predictor of the ratings produced by external coders and as a measure of discriminant validity we also investigated whether self-reported arousal was a significant predictor of the ratings. Note, a random intercept for Participant was included in these models. As hypothesised, for both datasets, self-reported valence was a significant predictor of the ratings based on video clips and self-reported arousal was not (see Table 2), suggesting that the manual coding conducted in this study does capture well the self-reported emotional valence experienced by the learners. Also as expected, the models did not explain much variance in valence ratings with Conditional R^2 values under 25%. In Study 1 the fixed effects explained 5.5% of variance (Marginal R^2) and

Table 3 Estimates for fixed effects in the best fitting LMMs for mean pen pressure in Study 1 and Study 2

Study 1				Study 2			
Predictor	Estimate	95% CI	*p*	Predictor	Estimate	95% CI	*p*
(Intercept)	0.20	0.19–0.20	<0.001	(Intercept)	0.21	0.20–0.21	<0.001
Control: Calibration mean pressure [a]	0.60	0.53–0.67	<0.001	Control: Calibration mean pressure [a]	0.66	0.60–0.71	<0.001
Accuracy of answer	−0.01	-0.01–0.00	0.106	Accuracy of answer	0.01	−0.0004 0.01	0.066
Phase $_{\text{Phase 2}}$	0.09	0.08–0.09	<0.001	Phase $_{\text{Phase 2}}$	0.09	0.09–0.10	<0.001
Phase $_{\text{Phase 3}}$	0.08	0.07–0.09	<0.001	Phase $_{\text{Phase 3}}$	0.09	0.09–0.10	<0.001
Item difficulty [a]	0.03	0.02–0.05	<0.001	Item difficulty [a]	−0.01	−0.03–0.01	0.469
Accuracy × Phase $_{\text{Phase 2}}$	0.01	0.005–0.02	0.002	Accuracy × Item difficulty [a]	0.03	0.01–0.05	0.001
Accuracy × Phase $_{\text{Phase 3}}$	0.03	0.02–0.04	<0.001	Accuracy × Phase $_{\text{Phase 2}}$	0.001	−0.01–0.01	0.789
Item difficulty [a] × Phase $_{\text{Phase 2}}$	0.03	0.01–0.05	0.002	Accuracy × Phase $_{\text{Phase 3}}$	0.01	0.01–0.02	<0.001
Item difficulty [a] × Phase $_{\text{Phase 3}}$	0.02	0.004–0.04	0.020	Item difficulty [a] × Phase $_{\text{Phase 2}}$	0.02	0.01–0.04	0.013
				Item difficulty [a] × Phase $_{\text{Phase 3}}$	0.03	0.01–0.05	0.001

p-values are based on the Wald statistic

[a] variable was mean centred before entering model

the full model (i.e., random and fixed effects) explained 21.1% of variance (Conditional R^2). In Study 2, Marginal $R^2 = 0.07$ and Conditional $R^2 = 0.15$.

What predicts mean pen pressure?

The best fitting LMMs for the mean pen pressure are presented in Table 3. Overall, the models explained a reasonable proportion of variance in mean pen pressure. The dataset for pen pressure (mean and variance) in Study 1 comprised 9522 observations from 243 participants (correctly answered items: 6813 trials from 242 participants; incorrectly answered items: 2709 trials from 217 participants). The fixed effects explained 40.5% of variance (Marginal R^2), the random effects (or, differences between individuals) explained an additional 20.8% of variance (Conditional $R^2 = 0.61$) and the Root Mean Square Error (RMSE[1]) was 0.068 (unit: proportion of maximum detectable pressure; range 0 to 1). Residual diagnostics indicated no violations of normality or homoscedasticity assumptions (see Supplementary Figure S1).

The dataset for pen pressure (mean and variance) in Study 2 comprised 12,627 observations from 341 participants (correctly answered items: 6522 trials from 324 participants; incorrectly answered items: 6105 trials from 336 participants). In the Study 2 model, the fixed effects explained 39.9% of variance (Marginal R^2), the random effects explained an additional 18.2% of variance (Conditional $R^2 = 0.58$), and the RMSE was 0.075. Residual diagnostics indicated no violations of normality or homoscedasticity assumptions (see Supplementary Figure S2).

In the following, the findings that are common across the two studies will be described. In both studies there is a positive association between the calibration mean pressure and the mean pressure used in the learning environment, suggesting some individual differences in pen pressure and highlighting the importance of controlling for these differences in analyses. In both studies, the main effect of Phase indicates that the mean pressure recorded increases across the phases. More specifically, more pressure is applied in Phases 2 and 3 compared to Phase 1; it should be noted that slide actions were required in Phases 2 and 3, whereas tapping on buttons was required in Phase 1. The Accuracy × Phase 3 interactions indicate that, contrary to our expectations, participants apply more pressure when moving on after positive feedback than negative feedback (in Phase 3). The Item difficulty × Phase interactions across both studies also indicate that in Phases 2 and 3, as item difficulty increases participants apply more pressure entering their answers and after reading their feedback, consistent with our hypotheses.

As well as these consistent findings there are some notable differences between the results of Study 1 and Study 2. In Study 1 only, higher item difficulty is associated with higher mean pen pressure also in Phase 1 (Typing Answer). In Study 2, as Item Difficulty increases, the effect of Accuracy on mean pressure increases (i.e., for harder items, more pressure is applied when they are answered correctly than incorrectly). Altogether, these results may indicate that mean pen pressure as measured here reflects arousal more than frustration.

[1] The RMSE from linear mixed-effects models quantifies the typical magnitude of residual prediction error, representing the average deviation between observed and model-predicted values on the outcome scale.

Table 4 Estimates for fixed effects in the best fitting LMMs for variance in pen pressure in Study 1 and Study 2

Study 1				Study 2			
Predictor	Estimate	95% CI	*p*	Predictor	Estimate	95% CI	*p*
(Intercept)	0.06	0.05–0.06	<0.001	(Intercept)	0.06	0.06–0.07	<0.001
Control: Calibration variance pressure [a]	0.46	0.40–0.52	<0.001	Control: Calibration variance pressure [a]	0.43	0.37–0.48	<0.001
Item difficulty [a]	0.01	0.01–0.02	<0.001	Accuracy of answer	−0.002	−0.004–0 001	0.220
Phase $_{\text{Phase 2}}$	0.05	0.05–0.05	<0.001	Item difficulty [a]	0.01	0.002–0.02	0.022
Phase $_{\text{Phase 3}}$	0.05	0.05–0.05	<0.001	Phase $_{\text{Phase 2}}$	0.05	0.04–0.05	<0.001
Accuracy × Phase $_{\text{Phase 1}}$	−0.002	−0.01–0.001	0.174	Phase $_{\text{Phase 3}}$	0.05	0.04–0.05	<0.001
Accuracy × Phase $_{\text{Phase 2}}$	0.002	−0.001–0.005	0.241	Accuracy × Phase $_{\text{Phase 2}}$	0.003	−0.001–0.01	0.097
Accuracy × Phase $_{\text{Phase 3}}$	0.005	0.002–0.01	0.001	Accuracy × Phase $_{\text{Phase 3}}$	0.01	0.004–0.01	<0.001
				Phase $_{\text{Phase 2}}$ × Item difficulty [a]	−0.01	−0.02–0.01	0.396
				Phase $_{\text{Phase 3}}$ × Item difficulty [a]	0.002	−0.01–0.02	0.830
				Accuracy × Item difficulty [a] × Phase $_{\text{Phase 1}}$	− 0.01	−0.03–0.001	0.064
				Accuracy × Item difficulty [a] × Phase $_{\text{Phase 2}}$	0.01	−0.01–0.02	0.204
				Accuracy × Item difficulty [a] × Phase $_{\text{Phase 3}}$	0.01	−0.01–0.02	0.238

p-values are based on the Wald statistic

[a] variable was mean centred before entering model

Table 5 Estimates for fixed effects in the best fitting LMM for facial expression valence rating in Study 1 and Study 2

Study 1				Study 2			
Predictor	Estimate	95% CI	*p*	Predictor	Estimate	95% CI	*p*
(Intercept)	2.49	2.42–2.57	<0.001	(Intercept)	2.55	2.50–2.60	<0.001
Accuracy of answer	0.61	0.53–0.69	<0.001	Accuracy of answer	0.74	0.68–0.80	<0.001
Accuracy $_{\text{Incorrect}}$ × Item difficulty [a]	0.36	0.04–0.67	0.027				
Accuracy $_{\text{Correct}}$ × Item difficulty [a]	−0.30	−0.49–−0.11	0.002				

p-values are based on the Wald statistic

[a] variable was mean centred before entering model

What predicts the variance of pen pressure?

The best fitting LMMs for the variance of pen pressure are presented in Table 4. These models explained less variance than those predicting the mean pen pressure: In Study 1 the fixed effects explained 37.6% of variance (Marginal R^2) and the random effects explained an additional 8.7% of variance (Conditional $R^2 = 0.46$); in Study 2 the fixed effects explained 27.4% of variance (Marginal R^2) and the random effects explained an additional 9.7% of variance (Conditional $R^2 = 0.37$). The RMSE was 0.035 for the Study 1 model and 0.039 for the Study 2 model. Residual diagnostics, which can be viewed in Supplementary Figures S3 and S4, indicated good model fit.

The following findings were common across the two studies. Similar to the findings for mean pressure values, the control variance variables calculated from the calibration trials are strong predictors of the variance measured within the learning environment, again pointing to the existence of stable individual differences in motor variability. Also mirroring the previous findings, the variance in the pressure applied increased across phases, with more variance in Phases 2 and 3 (slide actions) than in Phase 1 (tapping actions). More variable pressure is also applied as item difficulty increases, as expected. Lastly, more variance was measured in pen pressure when participants moved on after correct than incorrect feedback (Accuracy × Phase $_{\text{Phase 3}}$ interaction). Similar to the results of mean pen pressure, these results may suggest that variability in the pen pressure applied in our learning context reflects both arousal and effort.

In Study 2, an additional two-way and three-way interaction improved the model fit but did not themselves emerge as statistically significant predictors in the best fitting model. The only interaction that perhaps warrants attention is the marginally significant Accuracy × Item difficulty × Phase $_{\text{Phase 1}}$ interaction. The associated negative estimate indicates that in Phase 1 (Typing Answer), the difference in pen pressure variance between correct and incorrect answers becomes more negative as item difficulty increases. That is, for harder items, the difference in variance between correct and incorrect answers grows, with correct answers showing lower variance. This may indicate a lack of confidence when typing answers for difficult items that are in the end incorrect.

What predicts facial expression valence rating?

The best fitting LMMs for the facial expression valence rating (based on the video clips during feedback presentation; Phase 3) are presented in Table 5. The dataset for facial expression ratings in Study 1 comprised 2055 observations from 200 participants (correctly answered items: 1452 trials from 195 participants; incorrectly answered items: 603 trials from 176 participants). The dataset for facial expression ratings in Study 2 comprised 2991 observations from 317 participants (correctly answered items: 1611 trials from 289 participants; incorrectly answered items: 1380 trials from 298 participants). These models explained approximately one quarter of the variance in valence ratings. In Study 1 the fixed effects explained 12.4% of variance (Marginal R^2) and the random effects explained an additional 8.9% of variance (Conditional $R^2 = 0.21$); in Study 2 the fixed effects explained 16.1% of variance (Marginal R^2) and the random effects explained an additional 7.5% of variance (Conditional $R^2 = 0.24$). The RMSE was 0.669 for the Study 1 model and 0.785 (unit: facial expression valence rating; range 1 to 5) for the Study 2 model. Residual diagnostics, which can be viewed in Supplementary Figures S5 and S6, indicated acceptable model fit.

The consistent finding across the two studies is that children's facial expressions were rated with more positive valence when they had answered the quiz item correctly, namely when they were reading positive feedback (as was hypothesised). In Study 1, this additionally interacted with Item Difficulty, whereby in incorrectly responded trials, the more difficult an item was the more positive the facial expression was rated. Further, in correctly responded trials, the more difficult an item was the less positive the facial expression was rated. Possible interpretations of these findings will be elaborated upon in the Discussion section.

Discussion

The present study set out to evaluate the potential of two types of trace data—pen pressure and facial expressions—to provide meaningful insights into young learners' experiences during classroom-based self-study. Building on prior work that has investigated these indicators primarily with adults and/or under controlled conditions, we examined the feasibility of collecting them and their validity in authentic classroom contexts. Specifically, we investigated (1) whether mean pen pressure varied as a function of answer accuracy, item difficulty, and learning phase; (2) whether the variance of pen pressure was similarly predicted by these variables; and (3) whether facial expression valence ratings following feedback were influenced by answer accuracy and item difficulty. To this end, we analysed data from two separate studies with large samples of primary school children learning under natural conditions within a tablet-based digital learning environment. The studies demonstrated that valuable trace data can be collected unobtrusively from children without disturbing their natural learning processes using tablet devices available in many classrooms, and that these data can provide novel insights into the dynamic learning experiences inherent to self-study. The results common to both datasets highlighted stable aspects of these indicators, whereas study-specific findings highlighted the context dependency of these variables. These will be elaborated in the following paragraphs.

First, consistent with our expectations, participants entered their answers with more pressure as item difficulty increased. This finding is consistent with prior findings regarding cognitive load (Altmeyer et al., 2023) and expertise (Oviatt et al., 2018), but to our knowledge is the first demonstration that applied pen pressure varies as a function of item difficulty in primary school children. The chance to collect a proxy of children's experienced difficulty in an unobtrusive fashion may prove valuable for future studies as pen pressure could replace the collection of potentially disruptive trial-by-trial judgments from children.

Based on Schrader and Kalyuga's (2020) observation of a negative relationship between mean pen pressure and reported enjoyment, we expected that children would apply more pressure when processing negative than positive feedback. In fact, in our data participants applied more pressure when moving on after reading positive feedback than error feedback. This may reflect the heightened arousal elicited by positive feedback. In Schrader and Kalyuga's study, pressure was measured while students wrote newly learned Japanese characters on a tablet, and more pressure was associated with more frustration and less enjoyment. Putting these findings together, it seems that the phase in which pen pressure is measured is crucially important for its interpretation. Indeed, contrary to our hypotheses, there was no consistent evidence that mean pen pressure while typing or entering answers was associated with the accuracy of participants' answers.

There is also some evidence for context-specific findings regarding mean pen pressure, namely findings specific to only one study. In the data from Study 1, participants applied more pressure when typing their answers as item difficulty increased, as hypothesized. In Study 2, which had an overall higher difficulty, the effect of accuracy on mean pressure was amplified as item difficulty increased (across all phases). That is, for easier items there was little difference between correctly and incorrectly responded trials, but for harder items participants applied more pressure on items they ended up answering correctly than incorrectly. This could be interpreted in different ways, depending on which phase of the learning environment is considered. While typing and entering the answer, pressure may reflect effort and engagement, and difficult items that receive a lot of effort are more likely to lead to correct responses. If positive feedback heightens arousal, causing higher pressure after reading the feedback, this result could reflect especially high arousal when a participant succeeds on a particularly challenging item, indicating that children take item difficulty into account when appraising their success. Since this two-way interaction did not interact with Phase, both interpretations could be accepted.

Regarding research question 2 about variance of pen pressure, the findings largely mirrored the mean pen pressure findings. Consistent with previous findings (Altmeyer et al., 2023), participants' pen pressure was more variable as item difficulty increased. A novel finding was that pen pressure was more variable when participants moved on after receiving positive feedback than error feedback. This can also be interpreted as reflecting increased arousal. One additional result from the Study 2 dataset could be interpreted as reflecting metacognitive processes. That is, when typing answers on especially hard items that turn out to be incorrect, participants' pen pressure is more variable than when those answers turn out to be correct. This may indicate that children had insight into the difficulty of the items, and that increased pen pressure variance reflects their lack of confidence in the answer they were typing. This finding again highlights that the interpretation of an indicator, in this case variance of pen pressure, depends on the specific context in which it is measured (in this case, the phase within the trial and the overall difficulty level of the task).

In addition to addressing our research questions, the data collected in the current study also highlighted that there are considerable individual differences in the mean and variance of pen pressure applied, and that these (measured via the calibration baseline values) are strong predictors of the pen pressure values measured within the digital learning environment. As such, we would recommend including such calibration trials in studies that aim to analyse pen pressure. It is notable that the effect of item difficulty on pressure was replicated in our study, suggesting that this effect is rather stable and providing evidence that it can also be observed in younger participants engaged with a learning task in a classroom setting.

Research question 3 concerned children's facially expressed emotional valence as they worked on the learning task, as rated by external observers. Consistent with previous studies (Horvers et al., 2025; Merrick & Fyfe, 2023) and our hypothesis, participants' facial expressions were more positive following positive than error feedback. Study 1 provided some novel insights into the impact of individual item difficulty on facially expressed emotions. Specifically, when a more difficult item was answered incorrectly, participants expressed less negative valence than when an easier item was answered incorrectly. This likely indicates that the children identified the item as challenging and therefore did not experience such negative emotions when the error feedback was presented; that is, they may have attributed the failure externally (e.g., "It's not so bad, it was a really hard item"; Weiner, 1986) which served to minimise negative emotions. On the other hand, on correctly responded trials, harder items elicited less positive reactions than easier items. This finding may have emerged if children experienced a feeling of relief or surprise at having answered correctly, or if they had good metacognitive awareness of their success and therefore the feedback had limited impact. Based on the data collected in this study, we cannot distinguish between these different explanations. Nevertheless, these findings can be interpreted as reflecting some metacognitive processes, as children's assessment of item difficulty and insight into their own likelihood of success may result in greater or lesser expressed emotional valence.

Over and above the findings related to this research question, our ratings of facial expression valence did concur with children's self-reported emotional valence, consistent with previous findings (Horvers et al., 2025; Reisenzein et al., 2014). The fact that there was no association with self-reported arousal indicates that our observational rating scheme had high validity, and the high ICC between two raters demonstrates high reliability. Nevertheless, the low Conditional R^2 values also indicate that observed valence and self-reported valence are not identical and perhaps provide complementary information (as also concluded by Horvers et al., 2025). The data patterns confirm that children do show emotional reactions when receiving feedback on their performance, and social desirability or masking of negative emotions did not seem to present an issue here. Further, consistent with Greipl et al. (2021), we observed some dataset-specific findings, highlighting that facial expression valence ratings need to be interpreted with the specific context in mind.

Not only did we observe context-specific findings in each of the data streams (pen pressure and facial expression), but also data stream-specific findings. When considered together, we can see that each data stream delivers valuable and differentiated insights into children's learning experiences. Across both studies, after reading positive feedback children applied higher and more variable pen pressure, and expressed more positive valence in their faces, compared to negative feedback. While higher item difficulty was generally associated with increases in mean pen pressure and its variance, its association with facial expression valence was modified by whether the answer was correct or incorrect (in Study

1). The data from Study 2 also showed such an interaction of accuracy and item difficulty in pen pressure measures. A full comparison or integration of the insights gained from the two data streams is limited by the fact that facial expression was only recorded during the feedback phase, whereas pen pressure was measured across different learning phases. As such, pen pressure data and facial expression valence ratings were analysed separately, precluding examination of their relative explanatory power and temporal interplay. Future research could integrate these multimodal measures to investigate how implicit motor signals and emotional expressions converge or diverge during task performance (similar to Li et al., 2024 and Paans et al., 2019). Sequential analysis of these data streams would be particularly promising, as it could reveal whether changes in pen pressure precede shifts in facially expressed valence or vice versa. Such an approach would provide novel insights into the dynamic coupling of embodied and affective processes underlying self-regulated learning in young learners.

Limitations and future directions

The current study has some limitations that should be acknowledged. First, our method of calculating item difficulty was based on the sample performance and is as such sample-dependent. Alternative approaches include Item Response Theory, which defines difficulty as the ability level at which there is a 50% chance of success, and expert- or judgment-based estimates. Our data pattern may have been different if item difficulty had been calculated in a different way or if a greater range of difficulties had been intentionally implemented in the learning environment. Second, we measured pen pressure in three phases (typing answer, entering answer, after feedback) but crucially we did not measure pen pressure while participants worked on the maths items, as they used paper and pencil for this. It could be that pen pressure measured during the problem-solving phase would be even more closely linked to motivational processes and emotional experiences, and future studies should aim to collect this data as well. In our case, it was considered a trade-off with ecological validity, as making notes on a tablet would be rather novel and unusual for primary school children. Further, we were not able to implement handwritten responses, as children's handwriting could not be reliably identified by commercially available systems. Lastly, just as we observed that baseline pen pressure was a strong predictor of pen pressure during the learning activity, individuals probably differ in how much their underlying emotions are expressed on their faces. As such, efforts should be made to control this in studies that aim to use facial expression ratings to make individualised predictions.

Another task for future research is to investigate the validity of using automatic facial expression coding systems to analyse data collected from young children in real classroom settings. If this analysis method is established to be valid, it would broaden the contexts in which facial expressions could be evaluated and improve research efficiency. Although the current manuscript reports the analysis of two independent datasets and as such includes a form of replication, the two learning contexts were very similar. Future efforts should aim to uncover what sources of trace data are consistently informative across different school subjects, types of learning activities, and phases of learning. Lastly, we hope that the insights provided by the current correlational study can inspire and justify further studies into how different conditions influence children's learning experiences. Many of the post hoc explanations offered here deserve further targeted investigation, for example regarding children's

attributions and their effect on emotional reactions to errors, how children's metacognitive monitoring accuracy interacts with objective feedback, and whether effort invested in problem-solving is associated with pen pressure.

Conclusions

In conclusion, the current study provides initial evidence that pen pressure and facial expressions can serve as unobtrusive indicators of young learners' learning experiences in classroom-based self-study. While some effects replicated previous findings from studies with adults and/or conducted in laboratory settings, others highlighted the context-specific and individual-specific nature of trace data, underscoring the importance of careful interpretation and the collection of baseline measures. Future research should broaden the range of learning contexts and employ experimental approaches to establish when and how these indicators reliably reflect underlying cognitive, motivational, metacognitive, and emotional processes. By doing so, the field can move closer to identifying robust, scalable measures of self-regulated learning that are both ecologically valid and sensitive to the nuanced dynamics of children's classroom learning experiences.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-026-09482-0.

Acknowledgements We thank Valentina Zeitel and Robin Brandmeir for support with data collection, and Insha and Sahibjot Kaur for support with video coding. We are grateful to all participating children, teachers and schools for their contribution. This research was funded by a grant from the German Research Foundation (DFG) to Markus Dresel (DR 454/12 − 1) and Robert Grassinger (GR 5335/3 − 1).

Author contributions DB: Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, Review & Editing, Supervision. JS: Conceptualization, Methodology, Software, Investigation, Writing – Review & Editing, Project administration. CE: Investigation, Writing – Review & Editing. RG: Funding acquisition, Writing – Review & Editing. MD: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Funding acquisition.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets analysed and presented in this manuscript are freely available on OSF (https://osf.io/3apky).

Declarations

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

References

Altmeyer, K., Barz, M., Lauer, L., Peschel, M., Sonntag, D., Brünken, R., & Malone, S. (2023). Digital ink and differentiated subjective ratings for cognitive load measurement in middle childhood. *British Journal of Educational Psychology*, *93*, 368–385. https://doi.org/10.1111/bjep.12595

Azevedo, R., Bouchet, F., Duffy, M., Harley, J., Taub, M., Trevors, G., Cloude, E., Dever, D., Wiedbusch, M., Wortha, F., & Cerezo, R. (2022). Lessons learned and future directions of MetaTutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. *Frontiers in Psychology*, *13*, 813632. https://doi.org/10.3389/fpsyg.2022.813632

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Boekaerts, M. (2011). Emotions, emotion regulation, and self-regulation of learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of Self-regulation of Learning and Performance* (pp. 422–439). Routledge.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. R package version 1.1–7, URL http://CRAN.R-project.org/package=lme4

Boulahmel, A., Djelil, F., & Smits, G. (2025). Investigating self-regulated learning measurement based on trace data: A systematic literature review. *Technology Knowledge and Learning*, *30*(1), 119–156. https://doi.org/10.1007/s10758-025-09816-y

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Cole, P. M., Zahn-Waxler, C., & Smith, K. D. (1994). Expressive control during a disappointment: Variations related to preschoolers' behavior problems. *Developmental Psychology*, *30*(6), 835. https://doi.org/10.1037/0012-1649.30.6.835

Cole, P. M., Ram, N., & English, M. S. (2019). Toward a unifying model of self-regulation: A developmental approach. *Child Development Perspectives*, *13*(2), 91–96. https://doi.org/10.1111/cdep.12316

Conijn, J. M., Smits, N., & Hartman, E. E. (2020). Determining at what age children provide sound self-reports: An illustration of the validity-index approach. *Assessment*, *27*(7), 1604–1618. https://doi.org/10.1177/1073191119832655

Cross, M. P., Acevedo, A. M., & Hunter, J. F. (2023). A critique of automated approaches to code facial expressions: What do researchers need to know? *Affective Science*, *4*(3), 500–505. https://doi.org/10.1007/s42761-023-00195-0

Eberhart, J., Bryce, D., & Baker, S. T. (2024). Staying self-regulated in the classroom: The role of children's executive functions and situational factors. *British Journal of Educational Psychology*, *94*(3), 995–1010. https://doi.org/10.1111/bjep.12700

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, *1*(1), 3–14. https://doi.org/10.1016/j.edurev.2005.11.001

Greipl, S., Bernecker, K., & Ninaus, M. (2021). Facial and bodily expressions of emotional engagement: How dynamic measures reflect the use of game elements and subjective experience of emotions and effort. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CHI PLAY), 1–25.

Horvers, A., Molenaar, I., Bosse, T., & Lazonder, A. W. (2025). Emotional responses to feedback in adaptive learning technologies for early mathematics education. *Learning and Instruction*, *99*, 102192. https://doi.org/10.1016/j.learninstruc.2025.102192

Karlen, Y., Bäuerlein, K., & Brunner, S. (2024). Teachers' assessment of self-regulated learning: Linking professional competences, assessment practices, and judgment accuracy. *Social Psychology of Education*, *27*(2), 461–491. https://doi.org/10.1007/s11218-023-09845-4

Kovanovic, V., Azevedo, R., Gibson, D. C., & Ifenthaler, D. (Eds.). (2023). Unobtrusive observations of learning in digital environments: Examining behavior, cognition, emotion, metacognition and social processes using learning analytics. Springer Cham. https://doi.org/10.1007/978-3-031-30992-2

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Li, Q., Luximon, Y., Zhang, J., & Song, Y. (2024). Measuring and classifying students' cognitive load in pen-based mobile learning using handwriting, touch gestural and eye-tracking data. *British Journal of Educational Technology*, *55*(2), 625–653. https://doi.org/10.1111/bjet.13394

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

McCoy, D. C., Koepp, A. E., Jones, S. M., Bodrova, E., Leong, D. J., & Deaver, A. H. (2022). An observational approach for exploring variability in young children's regulation-related skills within classroom contexts. *Developmental Science*, *25*, e13250. https://doi.org/10.1111/desc.13250

Merrick, M., & Fyfe, E. R. (2023). Feelings on feedback: Children's emotional responses during mathematics problem solving. *Contemporary Educational Psychology*, *74*, 102209. https://doi.org/10.1016/j.cedpsych.2023.102209

Ninaus, M., Greipl, S., Kiili, K., Lindstedt, A., Huber, S., Klein, E., Karnath, H., & Moeller, K. (2019). Increased emotional engagement in game-based learning–A machine learning approach on facial emotion detection data. *Computers & Education*, *142*, 103641. https://doi.org/10.1016/j.compedu.2019.103641

Noldus Information Technology (2022). *FaceReader* (Version 9.0) [Computer-Software]. Noldus Information Technology. https://www.noldus.com/facereader

OECD. (2023). *PISA 2022 Results (Volume II): Learning During – and From – Disruption*. PISA, OECD Publishing. https://doi.org/10.1787/a97db61c-en

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.

Pickal, A.J., Spear, J., Bryce, D., Hallmen, T., Enste, C. S., Grassinger, R., André, E.,& Dresel, M. (2026). Multimodally assessed adaptive reactions to errors inprimary school students and their relation to knowledge acquisition. *Learningand Individual Differences*, *129*, 102927. https://doi.org/10.1016/j.lindif.2026.102927

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In P. R. Pintrich, M. Boekaerts & M. Zeidner (Eds.), *Handbook of Self-regulation* (pp. 451–502). Academic.

Paans, C., Molenaar, I., Segers, E., & Verhoeven, L. (2019). Temporal variation in children's self-regulated hypermedia learning. *Computers in Human Behavior*, *96*, 246–258. https://doi.org/10.1016/j.chb.2018.04.002

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Papenmeier, F. (2024). Tablet pen pointer pressure demo [Computer software]. Pavlovia. https://gitlab.pavlovia.org/frank.papenmeier/tablet-pen-pointer-pressure-demo

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, R., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pekrun, R., Vogl, E., Muis, K. R., & Sinatra, G. M. (2016). Measuring emotions during epistemic activities: The Epistemically-Related Emotion Scales. *Cognition and Emotion*, *31*(6), 1268–1276. https://doi.org/10.1080/02699931.2016.1204989

Potamianou, H., & Bryce, D. (2026). When can we and when do we adapt? Evidence that conflict adaptation can transcend contexts early in childhood. *Acta Psychologica*, *263*, 106356. https://doi.org/10.1016/j.actpsy.2026.106356

R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Renninger, K. A., Hidi, S., Krapp, A., & Renninger, A. (2014). *The role of interest in learning and development*. Psychology.

Rintala, A., Wampers, M., Lafit, G., Myin-Germeys, I., & Viechtbauer, W. (2023). Perceived disturbance and predictors thereof in studies using the experience sampling method. *Current Psychology*, *42*(8), 6287–6301. https://doi.org/10.1007/s12144-021-01974-3

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development*, *72*(5), 1394–1408. https://doi.org/10.1111/1467-8624.00355

Rovers, S. F., Clarebout, G., Savelberg, H. H., De Bruin, A. B., & van Merriënboer, J. J. (2019). Granularity matters: comparing different ways of measuring self-regulated learning. *Metacognition and Learning*, *14*(1), 1–19. https://doi.org/10.1007/s11409-019-09188-6

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. https://doi.org/10.1016/j.ijhcs.2019.102374

Seufert, T. (2026). Transforming self-regulated learning – Multimodal insights and future directions. *Educational Psychology Review*, *38*(1). https://doi.org/10.1007/s10648-026-10119-6

Somerville, M. P., & Whitebread, D. (2019). Emotion regulation and well-being in primary classrooms situated in low-socioeconomic communities. *British Journal of Educational Psychology*, *89*(4), 565–584. https://doi.org/10.1111/bjep.12222

Statistisches Bundesamt (2025). *Bevölkerung in Privathaushalten nach Migrationshintergrund Insgesamt*. Retrieved September 25th, 2025 from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Migration-Integration/

Reisenzein, R., Junge, M., Studtmann, M., & Huber, O. (2014). Observational approaches to the measurement of emotions. In L. Linnenbrink-Garcia & R. Pekrun (Eds.), *International Handbook of Emotions in Education* (pp. 580–606). Routledge.

Vandevelde, S., Van Keer, H., Schellings, G., & Van Hout-Wolters, B. (2015). Using think-aloud protocol analysis to gain in-depth insights into upper primary school children's self-regulated learning. *Learning and Individual Differences, 43*, 11-30. https://doi.org/10.1016/j.lindif.2015.08.02

Weiner, B. (1986). *An attributional theory of motivation and emotion*. Springler.

Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., Almeqdad, Q., & Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning*, *4*(1), 63–85.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, *45*(4), 267–276. https://doi.org/10.1080/00461520.2010.517150

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. *Handbook of Self-regulation* (pp. 13–39). Academic.

Authors and Affiliations

Donna Bryce[1] · Jana Spear[1] · Cara-Sophie Enste[2] · Robert Grassinger[2] · Markus Dresel[1]

✉ Donna Bryce
donna.bryce@uni-a.de

Jana Spear
jana.spear@uni-a.de

Cara-Sophie Enste
cara.enste@ph-weingarten.de

Robert Grassinger
grassinger@ph-weingarten.de

Markus Dresel
markus.dresel@uni-a.de

[1] Department of Psychology, University of Augsburg, Universitätsstr. 10, Augsburg 86135, Germany

[2] Department of Educational Psychology, University of Education Weingarten, St.-Longinus-Str. 9, Weingarten 88250, Germany