Bednorz, David; Sommerhoff, Daniel; Heinze, Aiso

Tracing learning in a digital learning environment. Combining process-object theories, individual learning trajectories, and machine learning to predict concept acquisition in mathematics

Unterrichtswissenschaft 54 (2026) 3, S. 261-290

Quellenangabe/ Reference:
Bednorz, David; Sommerhoff, Daniel; Heinze, Aiso: Tracing learning in a digital learning environment. Combining process-object theories, individual learning trajectories, and machine learning to predict concept acquisition in mathematics - In: Unterrichtswissenschaft 54 (2026) 3, S. 261-290 - URN: urn:nbn:de:0111-pedocs-361759 - DOI: 10.25656/01:36175; 10.1007/s42010-026-00245-z

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-361759
https://doi.org/10.25656/01:36175

Unterrichtswissenschaft (2026) 54:261–290
https://doi.org/10.1007/s42010-026-00245-z

Tracing learning in a digital learning environment: combining process-object theories, individual learning trajectories, and machine learning to predict concept acquisition in mathematics

David Bednorz · **Daniel Sommerhoff** · **Aiso Heinze**

Received: 17 February 2025 / Revised: 27 May 2026 / Accepted: 30 May 2026 / Published online: 17 June 2026

Abstract The acquisition of mathematical knowledge frequently poses challenges for students. It can be described in terms of Sfard's theory of reification, which encompasses the acquisition of operational and structural knowledge through the processes of interiorization, condensation, and reification. Applying this theory to mathematical concepts, such as the derivative, allows a theoretical account for identifying relevant learning phases in concept acquisition and identifying risks such as the development of pseudostructural knowledge, that is, structural knowledge without a sufficient operational foundation, which may hinder sustainable further learning Data from digital learning environments enable tracing individual learning trajectories during concept acquisition and identifying learners who may need additional support. This study examines whether individual learning trajectories derived from a digital learning environment implemented in real classrooms can support early, low-stakes identification of students at risk of developing pseudostructural knowledge of the derivative, and which theoretically derived learning phases are most predictive. Using data from 365 students from 15 upper-secondary classes, we aggregated students' performance into indicators aligned with the subconcepts of the derivative and Sfard's phases of concept acquisition. Machine-learning classifiers achieved modest but systematic discrimination based on these theory-grounded trajectory indicators. Adding covariates, including prior knowledge and cognitive abilities, yielded selective improvements and increased robustness under conservative operating rules, illustrating the recall–precision trade-off inherent in screening applications. Across models, early difference-quotient indicators, particularly those targeting condensation, contributed most to screening performance, whereas later indicators showed smaller incremental contributions. Overall, theory-based process indicators, comple-

✉ David Bednorz · Daniel Sommerhoff · Aiso Heinze
Department of Mathematics Education, IPN Leibniz Institute for Science and Mathematics Education, Olshausenstrasse 62, 24118 Kiel, Germany
E-Mail: bednorz@leibniz-ipn.de

mented by learner prerequisites, provide a substantively interpretable basis for low-stakes screening of learners at risk of acquiring pseudostructural knowledge of the derivative.

Keywords Derivative · Learning trajectories · Process-object-approaches · Concept acquisition · Digital learning environments

1 Introduction

Digital learning environments can provide structured learning opportunities while generating fine-grained process data (e.g., task attempts, responses, and interaction traces). In doing so, they can also make learners' increasing heterogeneity over time visible and support adaptive, individualized instruction (Schiller et al. 2024). To leverage such data, learning-analytics approaches increasingly use machine-learning methods to model learning outcomes and to identify learners who may benefit from timely, targeted support (e.g., Alnasyan et al. 2024; Faber et al. 2017; Wang et al. 2023). However, deriving actionable inferences from process data requires an explicit theoretical account of how conceptual understanding develops over time (Gobert et al. 2013). In particular, while process data do not directly capture knowledge acquisition, they enable the reconstruction of temporally ordered performance patterns that can be interpreted as indicators of learners' progress through instructional sequences in authentic classroom settings, provided that these indicators are aligned with a theory-based account of how conceptual understanding is expected to develop over time (e.g., Hakimi et al. 2021; Hillmayr et al. 2020; Lu et al. 2023).

In this study, we developed a theory-based digital learning environment for mathematics and used it to analyze students' learning trajectories. Sfard's process-object approach (1991, 1992) provides the theoretical framework: it conceptualizes mathematical concept acquisition as progressing through interiorization, condensation, and reification. Applied to the derivative, this framework is well suited because the derivative is foundational for further learning in mathematics and widely documented as challenging for many students (Berry and Nyman 2003; Kleiner 2001; Zandieh 2000). A key risk pattern in this framework is that learners may exhibit the appearance of structural knowledge that is not grounded in operational knowledge (pseudostructural knowledge), which may hinder sustainable concept development (Litteck et al. 2025; Sfard 1991). Against this theoretical background, we use students' process data to identify learners who may be at risk of developing pseudostructural knowledge of the derivative by the end of the unit. Because these temporally ordered process data capture increasing heterogeneity in learners' progress, they offer the opportunity to identify students early and in a low-stakes manner, enabling timely follow-up assessment and targeted support. Accordingly, we frame at-risk identification as an early screening task aimed at minimizing missed at-risk learners (prioritizing sensitivity/recall), while accepting a higher false-positive rate as a trade-off when the consequences of providing support to students who are not at risk are low and the classification can be refined through subsequent evidence. Using data from a theory-based digital learning environment on the derivative, this study investigates whether learners' fine-grained process data, interpreted through the phases of concept acquisition specified in Sfard's framework, can be used with machine-learning methods to identify learners at risk.

2 Digital learning environments in classrooms

Digital learning environments (DLE) can enhance learning and instruction, depending on their instructional design and implementation conditions. Meta-analyses show that technologies such as intelligent tutoring systems or dynamic visualizations can have particularly strong impacts on learning success (Hillmayr et al. 2020). DLE can support personalization through automated feedback and can adapt dynamically to learners' needs and ongoing performance, enabling instruction that is responsive to individual learning profiles (Chernikova et al. 2025; Hilz et al. 2025; Kucirkova et al. 2021; Plass and Pawar 2020; Tetzlaff et al. 2021). Furthermore, DLEs can automatically generate fine-grained, time-stamped log data on students' interactions (e.g., task attempts, responses). Whether such data provide meaningful evidence about learning depends on the alignment between task design and the targeted constructs (Gobert et al. 2013; Winne 2020). This alignment is particularly important when process data are used not only to score isolated responses but also to reconstruct learners' progress over time. Many systems rely mainly on selected-response formats (e.g., multiple-choice) in computer-based assessments (e.g., He and Cui 2025). Selected-response formats can yield valid evidence for many constructs. However, they often produce relatively sparse process information, frequently reducing learners' work to a small set of discrete responses, thereby limiting insight into strategies and intermediate reasoning when the goal is to trace learning processes in ways that are interpretable relative to a specific theory of concept development (Anghel et al. 2024). More generally, meaningful inference about learning processes depends on content-rich activities that are explicitly aligned with the targeted conceptual progression, implemented in authentic classroom settings, and on the formative use of DLEs that generate interpretable evidence of learners' progress over time (e.g., Kubsch et al. 2022).

Machine learning (ML) methods are increasingly applied to analyze data from students' learning within DLEs (e.g., Romero et al. 2008; Alonso-Fernández et al. 2019). ML methods are particularly effective for large-scale, heterogeneous data streams from multiple sources, as they can detect complex, potentially non-linear patterns that are difficult to capture with more classical modeling approaches (e.g., Wu 2021; Riestra-González et al. 2021). In learning analytics, ML is therefore widely used to predict learning outcomes and to inform decisions about personalization, support, and scaffolding (e.g., Sønderlund et al. 2019). Yet, whether such predictions translate into improved instruction depends on the validity and richness of the underlying evidence and on how predictions are enacted in practice (Asselman et al. 2023; Azizah et al. 2024; Lavelle-Hill et al. 2024). Considering the wide range of ML approaches, classification and regression are among the most frequently applied methods. Classification models, in particular, are widely used to assign learners to diagnostically relevant categories, for example when predicting dropout, identifying students at risk of low achievement, or detecting misconceptions, thereby potentially informing timely support (e.g., Asselman et al. 2023; Baker and Inventado 2014; Papamitsiou and Economides 2014; Romero and Ventura 2020).

The quality of these classifications is important, as the predictions can be inaccurate. Typical issues emerge with false-positive and false-negative classifications. Misclassifying learners can mean overlooking those in need of support or unnecessarily intervening with others (Eegdeman et al. 2023). Furthermore, research comparing the added value of ML approaches to classical methodological approaches suggests that more complex ML methods are not always significantly superior to classical statistical models regarding misclassification. Results indicate that the added value of ML over classical statistical methods is context-dependent. In some educational prediction settings, logistic or generalized linear regression performs comparably to, or even better than, more complex ML approaches with respect to predictive accuracy and error metrics (Cornell-Farrow and Garrard 2020; Lou and Colvin 2025). In other domains, such as health prediction with large and complex data sources, some ML models can outperform classical regression models, although the size and consistency of this advantage depend strongly on the data structure and modeling context (Bjerre et al. 2024; Iwagami et al. 2024).

Taken together, these studies emphasize that ML classifications must be treated with caution. They require rigorous validation and thoughtful integration in educational settings to ensure that they contribute positively to learning support and concept acquisition rather than providing no or only little benefit (see also Pham et al. 2025). These concerns are especially salient when using ML to identify at-risk learners, because the consequences of false classifications differ substantially across use cases, for example, between low-stakes screening followed by additional assessment and high-stakes situations that implement predictions as direct decision rules (e.g. college applications). In this context, it becomes essential to evaluate predictive models systematically and in a way that matches the intended use case. Common ML evaluation metrics include accuracy, precision, recall, F1, and ROC-AUC (Bowers et al. 2013; Foody 2023; Powers 2020; Saito and Rehmsmeier 2015). Crucially, these metrics imply different optimization targets and trade-offs. Depending on the intended use, models may be tuned to prioritize precision (reducing unnecessary follow-up for false positives) or recall/sensitivity (minimizing missed at-risk learners), often by selecting an appropriate decision threshold on predicted risk scores. ROC-AUC complements threshold-specific metrics by summarizing the model's ability to discriminate between at-risk and not-at-risk learners across thresholds.

In studies using classroom process data (e.g., Lam et al. 2024), predictive models hold promise, but their performance depends tightly on the representativeness and depth of the data and on choosing suitable evaluation metrics, which may vary depending on context and feature set (Dong et al. 2025). Depending on the intended use, different aspects of performance may matter more. For instance, in identifying students who need support, it is often more important not to miss anyone (high recall), while in other contexts, it might be more important to avoid unnecessary interventions (high precision) (e.g., Gašević et al. 2015).

To summarize, digital learning environments and ML approaches in educational contexts hold considerable potential, but they must be applied with careful attention to the theory-based and content-rich design of the environment, the specific predictive task being conducted, and the interpretation of model performance in light of the intended use. Therefore, for research on ML in education it is important to move beyond general claims of predictive potential and to examine which theory-based indicators combined with which type of ML provide actionable information at which point in an instructional trajectory. In the present study, we address this by linking ML-based screening to a theory-based learning trajectory.

3 Acquisition of mathematical concepts

Learning trajectories Learning trajectories (LTs) describe the acquisition of specific concepts over time as a sequence of steps, subconcepts, or components of what is to be learned (Clements and Sarama 2004; Confrey et al. 2014; Simon 1995; Sztajn et al. 2012). In mathematics education, the concept of LTs is closely linked to Martin Simon's depiction of a hypothetical learning trajectory (HLT) (Simon and Tzur 2004). A HLT specifies a content-specific sequence of conceptual steps that learners are expected to develop, together with learning activities and instructional supports designed to foster this development. In psychometric longitudinal research, by contrast, trajectories typically refer to empirically estimated growth curves based on repeated measures (Dumas and McNeish 2017; Mattison et al. 2022).

In contrast to the intended progression described by a HLT, learners' actual trajectories may deviate due to errors, misconceptions, difficulties, lack of motivation,

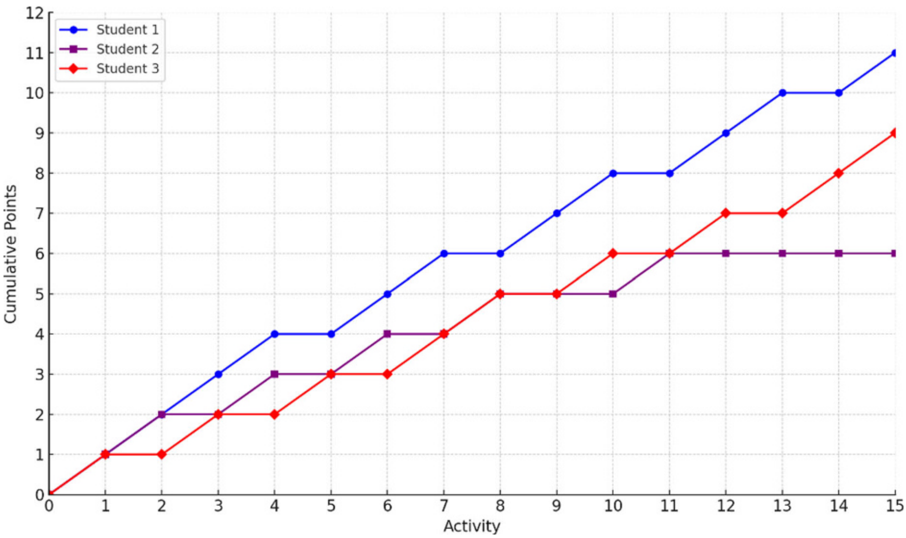

Fig. 1 *Schematic illustration of three learning trajectories (ILTs) across 15 activities. Note:* The plot shows cumulative task scores across activities as a descriptive visualization of attainment over the instructional sequence, primarily to illustrate progress and plateaus. It does not imply that activities are measured on a common psychometric scale, that scores can be combined into a joint scale, or that points are directly comparable across task formats and difficulty levels

and insufficient prior knowledge (Brodie 2014). Consequently, observed trajectories do not necessarily reflect the HLT (Clements and Sarama 2020; Confrey et al. 2014; Sarama et al. 2011). That's why we use the term individual learning trajectory (ILT) to denote a learner's progression through the conceptual steps targeted by the HLT, specifically, which steps are reached and how the learner moves between them over time. ILTs have been conceptualized in different ways, including domain-specific empirically validated progressions and data-driven reconstructions based on student performance (Barrett et al. 2012; Confrey et al. 2014; Sarama and Clements 2009; Sarama et al. 2011; Szilágyi et al. 2013).

This raises the question of how highly variable ILTs can be conceptualized. Because learning is only indirectly observable in classroom settings, ILTs can be operationalized through students' performance on sets of learning activities that are aligned with the intended HLT, that is, activities designed to target theoretically specified phase-subconcept steps in concept development within the digital learning

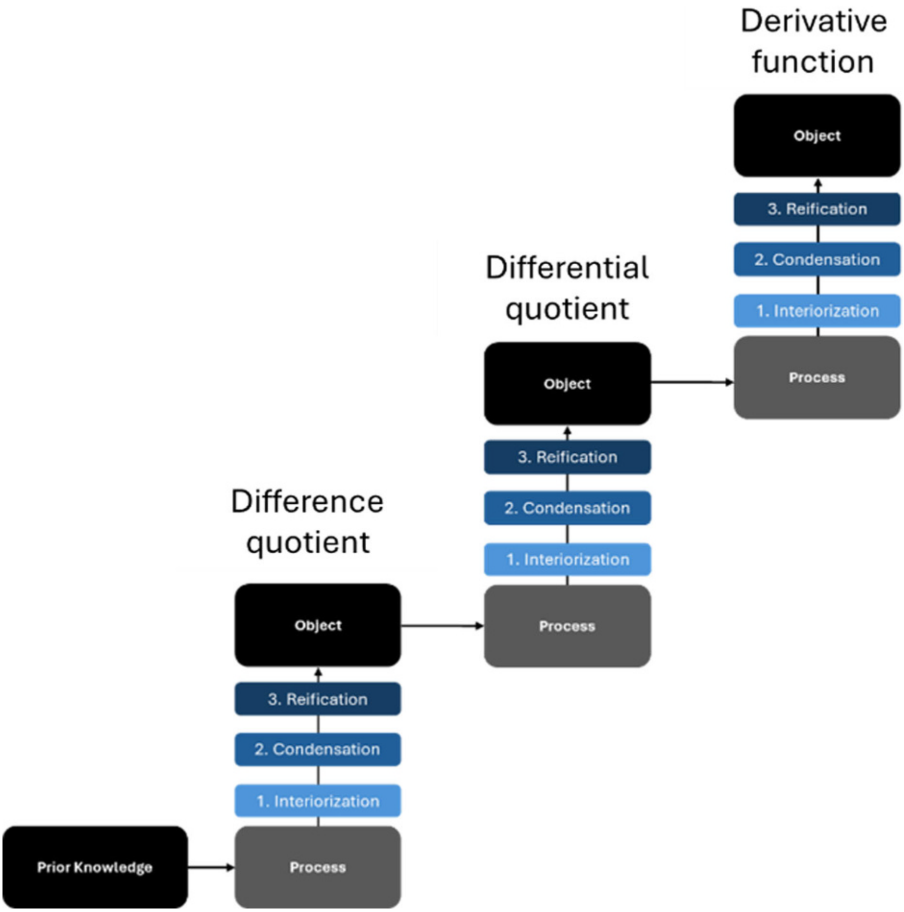

Fig. 2 Hierarchical structure of the concept of derivative according to the three subconcepts based on Sfard (1991)

environment (Mislevy and Haertel 2006). Specifically, students' aggregated success on learning activities aligned with the intended phase-subconcept structure can serve as an indicator of the degree to which students learned during the available instructional sequence (e.g., Faucon et al. 2020; He and Cui 2025). Figure 1 therefore provides a schematic illustration of three learners' trajectories across 15 activities within a lesson unit to visualize how performance patterns may unfold over an instructional sequence.

In the illustration, Student 1 and 3 show mostly steady progress, whereas Student 2 reaches a plateau from activity 11 onward. Such plateaus can indicate points in the instructional sequence where learners may require additional support.

Theories of concept formation in the case of the concept of derivative Sfard's theory of reification emphasizes the dual nature of mathematical concepts and describes concept acquisition as a hierarchical progression from operational to structural knowledge (Sfard 1991, 1992). Operational knowledge refers to perceiving a concept primarily as a process, whereas structural knowledge involves perceiving it as a unified object (Sfard 1991, 1992). These complementary perspectives provide a developmental account of how mathematical concepts are acquired, from process to object (Sfard and Linchevski 1994).

Within this framework, Sfard (1991, 1992) differentiates three phases of concept acquisition: interiorization, condensation, and reification. These phases characterize the transition from operational to structural knowledge. A further important aspect of concept acquisition based on Sfard's theory is that concept formation builds on "concrete objects," that is, previously available mathematical knowledge that enables learners to carry out processes in the first place (Sfard 1991, 1992). In addition, broader learner prerequisites such as domain-relevant prior knowledge and cognitive abilities can influence how effectively learners activate and use such prior mathematical objects (e.g., Bednorz et al. 2025; Hambrick and Engle 2002).

We apply this framework to the derivative to conceptualize a HLT (see Fig. 2). Following Zandieh (2000), we distinguish three subconcepts (difference quotient, differential quotient, and derivative function) each of which needs to be developed first operationally and then structurally. This structure guides the design of phase-specific learning activities in the learning environment (Appendix Table A1).

In view of the hierarchical order proposed in Sfard's framework and potential differences between a HLT and ILTs, it is also important to consider deviations from the intended HLT. Sfard highlights the problematic transition from operational to structural knowledge, where learners may come to operate with apparent structural elements without a solid operational foundation (so-called pseudostructural conception, Sfard 1992). Building on this, Litteck et al. (2025, p. 5) describe such structural knowledge that lacks operational grounding as *pseudostructural knowledge*. Learners may thus deviate from the HLT by acquiring knowledge components only superficially while at the same time appearing to operate with higher-level, but pseudostructural, components.

In summary, Sfard's hierarchical process-object approach differentiates operational and structural knowledge and provides a basis for specifying a derivative-related HLT (Sfard 1991, 1992; Sfard and Linchevski 1994). A central deviation from this trajectory is pseudostructural knowledge, apparent structural knowledge without sufficient operational knowledge, which has already been observed in the context of the derivative (Litteck et al. 2025).

4 Research questions

DLEs can provide structured learning opportunities and generate fine-grained process data on students' actions. ML methods can use such data to support early, low-stakes screening of learners who may require additional support. Building on Sfard's process-object approach and in line with Litteck et al. (2025), we conceptualize at-risk learners as students who are likely to develop pseudostructural knowledge by the end of the unit. Accordingly, we address the following research questions.

RQ1: To what extent can ML models based on students' ILTs identify learners at risk of developing pseudostructural knowledge, and at what point in the HLT do the predictions reach a practically acceptable level of performance for early detection?

We expected that learners' ILTs would provide predictive information about which students meet the end-of-unit at-risk criterion. More specifically, we anticipated that difficulties in activities targeting the difference quotient would provide an early indication of later risk, as insufficient operational grounding for structural knowledge of the difference quotient may foster the development of pseudostructural knowledge in subsequent subconcepts (cf. Litteck et al. 2025).

RQ2 extends the ILT-based screening perspective by testing whether adding learner prerequisites (domain-specific prior knowledge and cognitive abilities) improves at-risk classification beyond ILT indicators and whether learner prerequisites alone provide a viable baseline. Accordingly, we compare (a) ILT-only models across stages, (b) a prerequisites-only baseline, and (c) combined models that add prerequisites to stage-specific ILT indicators.

RQ2: How does at-risk classification performance compare across ILT-only, prerequisites-only, and combined (ILT + prerequisites) models across stages?

We expected that the at-risk classification would improve when learner prerequisites were added to the classification task. Although prior knowledge and cognitive abilities are theoretically important predictors of learning, we treat ILT-based screening as the practical reference point, because ILT process data are naturally generated during instruction with minimal additional burden, whereas collecting prerequisite measures (e.g., via prior-knowledge tests) typically requires dedicated assessment time and effort.

To examine RQ1 more closely and to combine learning analytics with mathematics education theory more specifically, we extend RQ1 by including in our analyses that process indicators are organized according to phases of concept acquisition (Sfard 1991, 1992) and subconcepts of the derivative (Zandieh 2000). Specifying

RQ1, we thus ask, which elements of the theoretical structure provide the most informative data for the identification task.

RQ3: Within the derivative learning trajectory, which phase-subconcept indicators contribute most to screening performance when predicting students' at-risk status?

There is limited empirical evidence on how Sfard's phases contribute to identifying learning and in particular at-risk learners across the subconcepts of the derivative. While reification and the condensation phase may be particularly informative, it remains unclear which phase-subconcept indicators provide the strongest screening information. To avoid a purely post-hoc reading of Sfard's framework, we specify two notions of "earliness": (a) earlier subconcepts in the instructional sequence (e.g., difference quotient) should yield earlier and stronger information than later subconcepts (e.g., derivative function), and (b) within each subconcept, interiorization and condensation should be especially informative as they establish operational knowledge that enables later reification. Evidence concentrated solely in the globally earliest phase indicators (e.g., those targeting the interiorization of the difference quotient) would provide limited support for phase-specific claims, whereas differentiated contributions across phases and subconcepts would better align with Sfard's theoretical assumptions.

5 Method

Sample The sample comprises 365 students from 15 classes from the first year of upper secondary level (grade 10 or 11, depending on the federal state). The gender distribution indicates that 52.3% of the students identify as female, 46.2% as male, and 1.5% as diverse. The learning environment comprises 16–20 lessons (of 45 min), depending on the pace at which the teacher covered the content, and requires corresponding digital resources in the schools. The study experienced high attrition (approx. 70% of participants) due to class-level dropout, with entire classes withdrawing (likely because of the study's complexity and length). Of the 15 classes that commenced, only five completed the entire study. At the conclusion of the unit, 108 students participated in the posttest.

Digital learning environment for the concept of derivative To trace individual learning trajectories, we designed a digital learning environment based on the assumed HLT of the concept of derivative (see Fig. 2). The environment operationalized the theory-derived HLT through sequential phase-subconcept activity sets, each comprising multiple activities across varied representations and contexts to provide sufficient opportunities for knowledge acquisition. All learners worked through the same standardized instructional sequence. The sequence was organized around the three derivative subconcepts: the difference quotient, the differential quotient, and the derivative function. Various digital activities based on STACK, GeoGebra, reactive and manipulable graphics and texts, and text inputs were used to design meaningful activities (see Appendix C for examples).

Instruments The study used a pre- (5 items) and a posttest (18 items), each focusing on the concept of derivative and which aligns with the intended learning goals of the unit, to measure students' learning gain throughout the unit for the concept of derivative. The tests were adapted from Litteck et al. (2025). Because the pretest items were also part of the posttest, both tests were scaled together using a one-dimensional Rasch model. Each item was additionally coded with respect to whether it required operational or structural knowledge (Litteck et al. 2025). The estimated person ability scores were used for further analysis. The reliability (EAP = 0.70; WLE = 0.65) obtained for the study proved to be satisfactory, underlining the reliability of the instruments used. The posttest results were used to identify learners at risk of pseudostructural knowledge, while the pretest results were included as a covariate representing domain-specific prior knowledge in the covariate-augmented model (RQ2). Cognitive abilities were assessed using a figural reasoning test (20 items, $\alpha = 0.91$) consisting of non-verbal pattern-completion and rule-induction items (matrix-/series-type figures), in which students select the missing element that best completes a visual pattern. Furthermore, prior knowledge relevant to learning the derivative was assessed with an IRT-scaled test (WLE reliability = 0.68), adapted from Litteck et al. (2025). The test covered prerequisite content from lower secondary mathematics (e.g., functions and graphs, slope/average rate of change, linear functions, and algebraic manipulation, but did not include concepts such as the limit), intended to capture domain-relevant knowledge for the instructional sequence. We used the posttest as a single end-of-unit outcome for the screening task. Although prior work (Litteck et al. 2025) suggests separable operational and structural knowledge facets of the concept of derivative, the posttest was not intended to support facet-level interpretations, nor does it mirror the HLT structure used to operationalize ILT information. We therefore applied a unidimensional scaling and did not define at-risk status in terms of separate operational and structural subscores, because the combined score showed acceptable measurement precision for the intended screening purpose, indicating that a unidimensional representation provides a parsimonious and sufficiently precise approximation in this dataset.

Definition of the dependent variable: "at-risk students" To define the at-risk students, we defined a content-based cutoff grounded in the IRT calibration of the posttest. Because of the scaling of the items, item difficulties (ξ) and person abilities (θ; WLEs) are situated on the same logit scale. Students whose ability estimates fall below a content-based difficulty level defined by key items can therefore be identified as at risk. In the IRT scaling, a person with ability θ solves an item with difficulty ξ with an expected success probability of 0.50 when $\theta = \xi$. We set the ability cut-off θ^* at 0.20 logits as a conservative minimum threshold for object-level attainment. Specifically, θ^* lies slightly above the upper difficulty bound of the lowest-difficulty object-level items ($\xi \leq 0.08$; Appendix Table A3), implying that learners at θ^* are expected to solve these boundary object-level items with $P \geq 0.50$ (and typically slightly above), while accounting for measurement uncertainty in WLE estimates ($M = 0.63$, SE range = 0.56–0.95 logits). Because the first process-level item starts at $\xi = 0.22$ (Appendix Table A3), learners below θ^* are even less likely to demonstrate process-level performance indicating operational knowledge.

Table 1 Definition of cumulative stages (ILT indicators included)

Stage	Phase-subconcept indicators
1	DC × INT
2	Stage 1 + DC × CON
3	Stage 2 + DC × REI
4	Stage 3 + DT × INT
5	Stage 4 + DT × CON
6	Stage 5 + DT × REI
7	Stage 6 + DF × INT
8	Stage 7 + DF × CON
9	Stage 8 + DF × REI

Note. DC difference quotient, *DT* differential quotient, *DF* derivative function, *INT* interiorization, *CON* condensation, *REI* reification. Each phase-subconcept indicator is a student-level mean success score (proportion correct) aggregated across activities mapped to that unit. Stages refer to cumulative predictor sets (Stage k includes all indicators available up to that point)

Table 2 Key Terms, Abbreviations and Their Meanings in This Study

Term	Meaning in this study
Hypothetical learning trajectory (HLT)	The theoretically specified learning sequence for acquiring the derivative concept (based on Sfard; structured by phases and subconcepts)
Individual learning trajectory (ILT)	The empirically observed learner-specific pathway through the instructional sequence, inferred from students' activity work
Phase (Sfard)	Conceptual phases of concept formation: interiorization, condensation, reification
Subconcept (derivative)	Difference quotient, differential quotient, derivative function
Phase-subconcept unit	Intersection of one phase and one subconcept; unit for aggregating activity-level performance into ILT indicators
Stage	Successive point in the instructional sequence at which cumulative evidence is available (stages accumulate phase-subconcept indicators observed up to that point)
At-risk learner (structural vs. operational)	Conception: elevated risk of ending the unit with acquiring pseudostructural knowledge. Operationalization: classification based on the pre-specified cut-off described in the Methods
Important Abbreviations	
DC, DT, DF	Difference quotient, Differential quotient, Derivative function
INT, CON, REI	Interiorization, Condensation, Reification

In line with Litteck et al. (2025), we therefore classify learners with $\theta < \theta^*$ as at risk of developing pseudostructural knowledge.

Although posttest performance is measured on a continuous scale, the practical target of the present study is an early screening decision. Modeling the binary at-risk criterion therefore aligns the statistical target directly with the intended decision. At the same time, we acknowledge that continuous-outcome prediction and two-step decision strategies (predict continuous performance and then apply the cut-off) are viable alternatives. We therefore evaluate these approaches in a robustness check (Appendix D).

Data analysis Students' work in the digital learning environment was automatically coded at the activity level as a parsimonious correctness indicator (correct/incorrect) and aggregated into a theory-driven 3×3 structure defined by derivative subconcepts (difference quotient, differential quotient, derivative function) and Sfard's phases (interiorization, condensation, reification), yielding nine ILT indicators per student. For each phase-subconcept unit, we computed a student-level mean success score (i.e., proportion correct) across all activities mapped to that unit. Aggregation across multiple activities within each phase-subconcept combination was used to reduce noise from single responses while preserving theoretically meaningful distinctions. To reflect the informational constraints of early identification, we evaluated cumulative feature sets aligned with the evidence available at successive points in the instructional sequence. We defined nine stages (k = 1 ... 9) such that Stage k included all ILT indicators available up to that point (i.e., Stage k = Stage k–1 plus the newly available phase-subconcept indicator), enabling prediction based on accumulated evidence rather than on a single indicator (see Table 1; for definitions of key terms and abbreviations, see Table 2).

Accordingly, the resulting stage indicators should be interpreted as summary evidence of phase-specific instructional-task success within the learning environment, rather than as psychometric scale scores. In line with this formative, theory-driven operationalization, the primary validity argument for using these indicators rests on (a) their explicit alignment with the phase-subconcept structure specified in the HLT and (b) their utility for predicting the end-of-unit outcome in a screening-oriented setting.

To handle missing data in covariates, we performed multiple imputation using the mice package (van Buuren and Groothuis-Oudshoorn 2011) in R, applying predictive mean matching. Diagnostic checks confirmed that the imputations preserved the data structure, and complete-case as well as sensitivity analyses yielded consistent results (see appendix Table A2).

Subsequent analyses were conducted in Python using scikit-learn (Pedregosa et al. 2011). Predictive performance was estimated via repeated cross-validation to obtain robust out-of-sample metrics and an empirical estimate of performance variability. Specifically, we used repeated 5-fold cross-validation with five repetitions for both regression and classification models, resulting in 25 test folds per model specification. All preprocessing steps (imputation, scaling, and model fitting) were performed strictly within the training folds to avoid information leakage.

To address RQ1 and RQ2, we estimated (a) models that predict learners' posttest outcomes and (b) models that identify learners classified as at risk. For classification, we trained six families of classifiers commonly used in educational prediction settings: logistic regression, support vector machines with an RBF kernel, random forest, gradient boosting, LightGBM, and XGBoost. We included logistic regression as a parsimonious, classical baseline model and compared all models under the same cross-validation and evaluation protocol. Model specifications were held constant across feature sets to enable comparability across stages and phases. Missing values in predictors were handled via median imputation within the training folds. Continuous predictors were standardized when required by the estimator (e.g., logistic regression, SVM). Because the intended use case is early, low-stakes screening, aimed at minimizing missed at-risk learners who may benefit from follow-up support, we prioritize recall (sensitivity) while explicitly constraining the rate of false positives. For internal model comparison, we therefore rely on the *F2*-score, which weights recall more strongly than precision while still penalizing overly liberal classification. To complement this recall-oriented summary with a decision-focused perspective, we additionally report screening performance at an operating rule defined by a minimum precision target ($P \geq 0.30$). In practical terms, this corresponds to accepting that up to 70% of identified learners may be false positives, a trade-off that is consistent with broad screening applications where identified cases are subject to subsequent assessment rather than immediate instructional action. This evaluation strategy aligns with established practices in educational early warning and learning analytics research (Baker and Inventado 2014; Hlosta et al. 2017; Kizilcec et al. 2017). Operating rules were determined fold-wise by scanning classification thresholds on the validation data within each cross-validation fold. If the precision floor could be achieved, we selected the threshold that maximized recall subject to $P \geq 0.30$. If no threshold met the precision floor (e.g., due to limited class separability and class imbalance), we transparently report that the target could not be satisfied. For completeness, we then report recall and precision at the unconstrained threshold that maximized recall and provide the corresponding precision-recall trade-off. This approach makes explicit when strict precision control cannot be guaranteed by the model and prevents optimistic reporting.

To address RQ3, we estimated feature importance using permutation importance (Breiman 2001; Fisher et al. 2019). For each stage, we selected the best-performing model according to the recall-sensitive evaluation described above and computed permutation importance under cross-validation. Within each split, the model was fitted on the training data, and the decrease in the evaluation metric on the held-out fold was recorded after permuting each predictor. Importance values were then averaged across splits to stabilize estimates. Importances were computed using the same fold-wise operating-point procedure (target $P \geq 0.30$ with the predefined fallback when unattainable), so that each importance value reflects a feature's contribution to maintaining recall under a practically motivated target on false alarms.

To quantify which stage indicators contributed most to identifying at-risk learners, we computed permutation importance for the best-performing classifier within each stage. Importance was evaluated using the same screening-relevant criterion as in the main classification analyses, i.e., recall at the defined precision target, ensuring that importance values reflect contributions to sensitivity under controlled false-alarm rates rather than to generic discrimination alone.

Permutation importance was computed as follows: For each trained model and evaluation split, we first computed the baseline performance M (recall under the precision target) using unpermuted predictors. We then permuted each predictor x_j in turn, recomputed the performance $M^{(j)}$, and defined the raw importance as the absolute performance loss: $\Delta_j = M - M^{(j)}$. To reduce over-interpretation of trivial fluctuations, we treated very small absolute changes as negligible and set them to zero (near-zero threshold: $|\Delta_j| < 0.005$ on the metric scale). Negative changes (i.e., apparent improvements after permutation) were also set to zero, as they typically indicate sampling noise rather than meaningful predictive structure.

For comparability, we report $\Delta\% = 100 \times (\Delta_j/M)$, where M denotes the fold's baseline recall under the precision target ($P \geq 0.30$) prior to permutation. Importance values were averaged across cross-validation repetitions (splits) to stabilize estimates. Entries with "—" in importance summaries indicate that either (a) a predictor was not part of the feature set available up to that stage, or (b) the predictor did not yield any non-negligible positive loss after applying the near-zero rule. These cases are distinguished in the table notes. Permutation importance values are conditional on the cumulative predictor set at each prediction stage. Thus, they quantify each indicator's marginal contribution beyond the indicators already included, rather than the importance of the prediction stage itself.

Robustness checks: alternative operationalizations of posttest performance To examine whether our conclusions depended on how posttest performance was operationalized, we conducted three complementary robustness analyses using the same cumulative feature sets and repeated cross-validation procedure as in the main analyses: (a) continuous prediction of posttest WLE, (b) a two-step predict—then—threshold strategy based on the same IRT-defined cut-off, and (c) a sensitivity analysis using a slightly shifted threshold. Across these checks, the substantive pattern of results remained stable (for details see Appendix D). Continuous prediction of posttest WLE explained only a small proportion of variance overall, and the two-step approach showed no systematic advantage over direct binary classification. Likewise, shifting the at-risk threshold slightly did not alter the main conclusions. Taken together, these robustness checks support our decision to focus the main analyses on direct binary classification with a substantively grounded at-risk definition, which provides the most parsimonious fit to the present early, low-stakes screening use case.

6 Results

RQ 1 & RQ 2: classifications of students at risk of acquiring pseudostructural

Table 3 Model performance for ILT-based models, summarized by stage. Metrics refer to the top-performing classifier per stage (by F2)

Stage	Model	F2	Recall	Precision	ROC-AUC	PR-AUC
1	SVM RBF	0.57	0.88	0.25	0.63	0.25
2	XGB	0.57	0.94	0.23	0.63	0.28
3	XGB	0.57	0.93	0.24	0.64	0.28
4	XGB	0.53	0.92	0.21	0.55	0.21
5	XGB	0.55	0.90	0.23	0.61	0.24
6	RF	0.55	0.91	0.22	0.61	0.24
7	XGB	0.58	0.92	0.25	0.64	0.29
8	XGB	0.58	0.90	0.25	0.64	0.28
9	XGB	0.57	0.92	0.23	0.63	0.27

Note. Models (including LogReg as a classical baseline): XGB, LGBM, SVM RBF, RF, GB

Table 4 ILT-based model (ILT only): Recall and precision at the selected operating rule (intended $P \geq 0.30$). Best per stage.

Stage	Model	Target met ($P \geq 0.30$)	Recall	Precision	Threshold
1	GB	Yes	0.46	0.30	0.241
2	GB	Yes	0.40	0.30	0.297
3	SVM RBF	Yes	0.47	0.30	0.202
4	LogReg	No	0.63	0.20	0.496
5	XGB	Yes	0.29	0.30	0.375
6	RF	Yes	0.43	0.30	0.478
7	RF	No	0.57	0.26	0.380
8	LGBM	Yes	0.45	0.30	0.283
9	GB	Yes	0.39	0.30	0.252

Note. Models (including LogReg as a classical baseline): XGB, LGBM, SVM RBF, RF, GB. Intended operating rule: maximize recall subject to precision ≥ 0.30. If no threshold achieved precision ≥ 0.30, we report the unconstrained threshold maximizing recall and mark the stage as 'Target not met'. Predicted-positive rates corresponding to the operating points are provided in Appendix B

knowledge on students' ILTs without and with covariates To answer Research Questions 1 and 2, we evaluated predictive performance across nine cumulative stages that correspond to successive points in the hypothesized learning trajectory. At each stage, models were trained on the information accumulated up to that point (i.e., all ILT indicators available up to the respective stage). Thus, differences across stages reflect how predictive performance changes as additional evidence becomes available during instruction, rather than effects of a single stage indicator. Because the practical use case is early, low-stakes screening, we focus primarily on whether the models are able to identify as many at-risk learners as possible, while also reporting complementary indices that characterize false-alarm rates and overall discrimination (precision, F2, PR-AUC, and ROC-AUC). Full model comparisons

Table 5 Model performance for covariate-augmented models, summarized by stage. Metrics refer to the top-performing classifier per stage (by F2)

Stage	Model	F2	Recall	Precision	ROC-AUC	PR-AUC
0	LogReg	0.53	0.95	0.20	0.59	0.23
1	LogReg	0.55	0.95	0.20	0.60	0.25
2	LogReg	0.54	0.96	0.20	0.60	0.24
3	LogReg	0.54	0.96	0.20	0.60	0.24
4	LGBM	0.51	0.91	0.20	0.51	0.22
5	LogReg	0.55	0.96	0.20	0.60	0.24
6	LogReg	0.55	0.96	0.20	0.60	0.26
7	RF	0.54	0.96	0.21	0.61	0.27
8	RF	0.54	0.94	0.21	0.62	0.28
9	RF	0.54	0.94	0.21	0.61	0.28

Note. Models (including LogReg as a classical baseline): XGB, LGBM, SVM RBF, RF, GB. Prerequisites only denotes a baseline model using only prior knowledge and cognitive ability measures, without any learning-stage (ILT) indicators. Stage 0 = Prerequisites only across classifiers are provided in Appendix B. The main text highlights the most informative patterns across stages.

When using ILT features alone (Table 3), models achieved high recall across all stages (ranging from 0.88 to 0.94), but showed only limited discrimination between at-risk and not-at-risk students. Accordingly, the high recall values reflect the fact that the models were optimized to miss as few at-risk learners as possible, whereas ROC-AUC and PR-AUC indicate that the overall separation between at-risk and not-at-risk learners remained only moderate. ROC-AUC values ranged from 0.55 to 0.64, and PR-AUC values remained low overall (between 0.21 and 0.29).

Across stages, the best-performing models reached F2-scores between 0.53 and 0.58, with XGB most frequently emerging as the top-performing classifier. In practical terms, this means that identifying most at-risk learners required thresholds that also classified many not-at-risk learners as at risk.

The frequent selection of XGB across stages further suggests that tree-based ensemble methods were often better able to capture nonlinear relationships among ILT features than the alternative classifiers.

We evaluated classification performance under an intended minimum precision target ($P \geq 0.30$). For the ILT-only models, this target was achievable in most, but not all, stages (Table 4).

More specifically, seven of the nine stages met the intended precision threshold, whereas Stages 4 and 7 did not and were therefore marked as target not met. At the selected operating rule, recall ranged from 0.29 to 0.63 across stages. The highest recall values were observed at Stage 4 (0.63), Stage 7 (0.57), and Stage 3 (0.47), but the precision target was not met in Stages 4 and 7. Restricting attention to stages that satisfied the intended precision level, recall ranged from 0.29 to 0.47, indicating that the models became considerably less sensitive once a stricter requirement on false alarms was imposed. Thus, compared with the recall-oriented F2 operating point,

Table 6 COV-augmented model: Recall and precision at the selected operating rule (intended $P \geq 0.30$). Best per stage.

Stage	Model	Target met ($P \geq 0.30$)	Recall	Precision	Threshold
0	LogReg	No	0.18	0.16	0.605
1	RF	Yes	0.48	0.30	0.340
2	RF	Yes	0.42	0.30	0.361
3	SVM RBF	Yes	0.43	0.30	0.195
4	LogReg	No	0.49	0.28	0.560
5	LogReg	Yes	0.35	0.30	0.632
6	LogReg	Yes	0.27	0.31	0.668
7	LogReg	Yes	0.35	0.30	0.641
8	GB	Yes	0.28	0.30	0.373
9	RF	Yes	0.39	0.30	0.373

Note. Models (including LogReg as a classical baseline): XGB, LGBM, SVM RBF, RF, GB. Under the fixed precision floor ($P \geq 0.30$), the target was met in eight of the ten stages, whereas Stages 0 and 4 were marked as target not met. No stage achieved the predefined recall target of 0.70. Stage 0 = Prerequisites only. Predicted-positive rates corresponding to the operating points are provided in Appendix B

Table 7 Permutation importance of ILT indicators when predicting at-risk status based on cumulative performance up to each stage (ILT model only)

	Prediction stage									
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8	Stage 9	Mean
Feature										
DC INT	0	80.35	65.33	13.68	7.51	6.56	2.42	1.26	2.00	18.68
DC CON	—	77.02	36.07	32.74	23.51	22.00	21.79	17.05	19.51	31.21
DC REI	—	—	28.39	8.84	10.70	12.32	5.72	6.67	8.42	11.58
DT INT	—	—	—	13.47	4.45	6.39	0.90	0.46	0.20	4.31
DT CON	—	—	—	—	0	0.11	0	0	0	0.02
DT REI	—	—	—	—	—	0	0	0	0	0
DF INT	—	—	—	—	—	—	0	0	**0**	0
DF CON	—	—	—	—	—	—	—	0	0	0
DF REI	—	—	—	—	—	—	—	—	0	0

Note. Permutation importance is reported as ΔRecall [%], the percentage decrease in recall after permuting the respective feature, computed separately for each prediction stage at the operating rule selected to meet (or, if not attainable, approach) the precision target ($P \geq 0.30$).—indicates that the feature was not included at the respective stage; 0 indicates that the feature was included but yielded no non-negligible loss after applying the near-zero rule.

the precision-floor operating point yields a more conservative but less sensitive screening rule.

Including covariates slightly improved model quality and stability across stages (Table 5). Logistic regression and random forest most frequently emerged as the top-performing classifiers, suggesting that the added covariates provided robust predictive information across stages. Importantly, performance improved slightly from the prerequisites-only baseline (Stage 0; F2 = 0.53, recall = 0.95, PR-AUC = 0.23) to Stage 1, when the first ILT indicator was added (F2 = 0.55, recall = 0.95, PR-AUC = 0.25), suggesting that early ILT evidence provides some incremental value beyond prerequisites, although the gain in global performance metrics was modest. Across subsequent stages, overall discrimination remained moderate and fairly stable (ROC-AUC = 0.51–0.62; PR-AUC = 0.22–0.28), and F2 scores stayed within a relatively narrow range (0.51–0.55). Recall remained high across all stages (0.91–0.96), while precision remained low (0.20–0.21), indicating that the covariate-augmented models also primarily favored sensitivity over specificity at the F2-optimized operating point.

At the operating rule based on the intended minimum precision target ($P \geq 0.30$), performance was more differentiated for the covariate-augmented models (Table 6). The target was met in most, but not all, stages. More specifically, the precision threshold was achieved in eight of the ten stages, including several later stages, whereas only Stages 0 and 4 were marked as target not met. This pattern suggests that, within the covariate-augmented specification, the predefined precision requirement could be satisfied in most stages. Compared with the ILT-only models across the shared instructional sequence (Stages 1–9), however, this did not increase the overall number of stages meeting the target, but rather changed the stage-specific pattern in which the precision criterion was satisfied. At the same time, the results indicate that the trade-off between detecting many at-risk learners and limiting false alarms remained substantial, as meeting the precision target still required a more conservative operating point than under F2-based optimization.

Comparing ILT-only and covariate-augmented models revealed modest, stage-dependent differences. Under F2 optimization, both specifications achieved very high recall. Overall discrimination was broadly similar, although covariate augmentation yielded selective improvements in some stages. In addition, the covariate-augmented models more often favored simpler classifiers, with logistic regression dominating in earlier stages and random forest in later stages, whereas the ILT-only models showed no equally consistent pattern across algorithms. Under the precision-floor target ($P \geq 0.30$), covariate augmentation did not increase the overall number of stages meeting the target across the shared instructional sequence (Stages 1–9: 7 of 9 in both specifications), but it changed the stage-specific pattern in which the precision criterion was satisfied. However, this did not remove the general trade-off between sensitivity and precision, which remained evident throughout the instructional sequence.

RQ 3 feature importance of phases and subconcepts Across all ILT-only models, permutation importance analyses indicate that screening-relevant predictive information is not concentrated in a single indicator, but distributed across multiple stages when predicting at successive cumulative stages of the trajectory (Table 7).

Table 8 Importance of feature (Δ Recall [%]) by phase and sub-concept (covariate-augmented model)

	DC INT	DC CON	DC REI	DT INT	DT CON	DT REI	DF INT	DF CON	DF REI	Mean
Feature										
DC (INT)	0.74	25.75	0.60	0	0	1.90	9.12	6.25	4.35	5.41
DC (CON)	—	37.79	0.25	0.28	0	0	23.12	20.98	21.09	12.94
DC (REI)	—	—	0	0	0	0	20.28	9.51	8.14	5.42
DT (INT)	—	—	—	0	0	0.20	15.79	10.81	6.11	5.49
DT (CON)	—	—	—	—	0	0	10.98	5.40	3.37	3.95
DT (REI)	—	—	—	—	—	0	0.28	0	0	0.07
DF (INT)	—	—	—	—	—	—	7.44	6.28	2.74	5.49
DF (CON)	—	—	—	—	—	—	—	7.02	0	3.51
DF (REI)	—	—	—	—	—	—	—	—	3.51	3.51
Covariates										
Domain-specific prior knowledge	8.89	31.75	0.14	0.14	0	1.20	18.49	10.00	6.49	8.57
Prior knowledge	1.12	47.02	0	0	0	0.50	17.89	15.44	14.74	10.75
Cognitive abilities	16.28	30.88	5.54	1.19	2.88	0	11.89	5.23	4.88	8.75

Note. Permutation importance is reported as ΔRecall [%] (percentage decrease in recall after permuting the respective feature), computed at the precision-constrained operating rule (target $P \geq 0.30$). Columns correspond to the cumulative prediction stage at which the respective phase-subconcept indicator becomes available;—indicates the feature was not included at that point

To avoid ambiguity, we distinguish two analytic levels here: (a) prediction stages, which represent cumulative time points in the instructional sequence (Stages 1–9), and (b) ILT indicators, which denote the phase-subconcept performance aggregates (DC × INT, ..., DF × REI) used as predictors within each stage. Mean decreases in recall after permutation ranged from near zero to 80.35%, reflecting substantial heterogeneity in the contribution of individual ILT indicators. Because all models are defined cumulatively, permutation importance values are conditional on the full set of predictors available at each prediction stage. Accordingly, a decreasing importance of individual early-stage indicators across later stages does not imply declining relevance. Instead, it reflects the increasing redundancy and shared predictive information among connected predictors as additional indicators are added.

Overall, the observed pattern suggests that most screening-relevant information is already captured when predicting based on performance accumulated up to the early and mid stages of the trajectory (Stages 1 to 5), whereas later indicators contribute comparatively little additional, non-redundant predictive information. Notably, a small subset of ILT indicators exhibits comparatively high permutation importance when predicting at-risk status across successive prediction stages, with particularly pronounced effects early in the sequence that attenuate as additional predictors are added. In contrast, most other indicators show persistently smaller marginal contributions once earlier information is accounted for.

Within the difference-quotient block (prediction Stages 1–3), the DC × INT indicator remains an important predictor of at-risk status even after adding the subsequent difference-quotient indicators (DC × CON and DC × REI). Importantly, the outcome is end-of-unit at-risk classification rather than later-stage performance. Across all prediction stages, DC × CON showed the highest average permutation importance ($M = 31.21\%$), followed by DC × INT ($M = 18.68\%$) and DC × REI ($M = 11.58\%$). Additional, though smaller, contributions occured after the introduction of the first differential-quotient indicator (DT × INT; $M = 4.31\%$). Taken together, these results suggest that most screening-relevant predictive information is already available when models use performance accumulated up to prediction Stage 4, whereas later indicators add comparatively little non-redundant information for distinguishing at-risk from not-at-risk learners.

When covariates were added (Table 8), their contribution to prediction varied across prediction stages. Permutation importances indicate that predictive contributions were again concentrated in early difference-quotient indicators (especially DC × CON and DC × INT), whereas the differential-quotient indicators (DT × INT/CON/REI) often yielded near-zero marginal contributions once included, particularly at later prediction stages. Compared with the ILT-only specification, mean permutation importances were slightly lower and more widely distributed, including indicators associated with the derivative function. This pattern suggests that trajectory-based evidence and learner prerequisites jointly inform screening performance in the covariate-augmented specification.

Among learner prerequisites, domain-specific prior knowledge and cognitive abilities contributed substantial predictive information when included ($M = 10.75$ and $M = 8.75$, respectively). In the covariate-augmented specification, DC × CON was the most informative ILT-based indicator ($M = 12.94\%$), while domain-specific prior

Table 9 Top 8 predictors for identifying at-risk learners across model specifications

Rank	Predictor	Model type	Mean Δ Recall [%]
1	DC (CON)	ILT-only	31.21
2	DC (INT)	ILT-only	18.68
3	DC (REI)	ILT-only	11.58
4	Prior Knowledge	Covariate-augmented	10.75
5	Cognitive abilities	Covariate-augmented	8.75
6	Domain-specific prior knowledge	Covariate-augmented	8.57
7	DT (INT)	Covariate-augmented	5.49
8	DT (INT)	ILT-only	4.31

Note. Predictors are ranked by mean permutation importance (percentage decrease in recall at the precision-constrained operating rule, target $P \geq 0.30$), averaged across prediction stages in which the predictor was included. ILT-only and covariate-augmented specifications are reported jointly to facilitate comparison. Full stage-wise importance values are provided in Tables 7 and 8

knowledge was the strongest learner-prerequisite predictor ($M = 10.75$), followed by cognitive abilities ($M = 8.75$). Overall, predictive information was concentrated in the early difference-quotient indicators (introduced in prediction Stages 1–3), whereas indicators introduced later in the instructional sequence contributed only modest and more intermittent additional information, including some derivative-function indicators (introduced in Stages 7–9). Outside these stages, marginal importances tended to be small, consistent with increasing redundancy once early differences are accounted for.

A more differentiated pattern emerges across cumulative prediction stages. Prior knowledge contributes notably when prediction is based on data up to Stage 2, but its importance decreases once the Stage-3 ILT indicator (DC × REI) is included. This suggests that the Stage-3 evidence captures part of the predictive information initially associated with prior knowledge. A complementary pattern is observed later in the unit. When prediction is based on data up to Stage 7, prior knowledge regains importance, and derivative-function ILT indicators (DF × INT/CON/REI) contribute more in the covariate-augmented specification than in the ILT-only specification. This suggests that derivative-function indicators contribute incremental predictive information mainly in interaction with learners' prerequisites, rather than providing strong standalone information, which is consistent with the view that advanced concept development builds on, but does not fully subsume, earlier knowledge.

Across specifications, permutation importances indicate that predictive information is concentrated in the early difference-quotient ILT indicators, whereas indicators introduced later contribute comparatively little additional information (Tables 7 and 8). In the ILT-only specification, the largest contributions stem from the difference-quotient condensation and reification indicators, with smaller contributions from early differential-quotient indicators (Table 7).

In the covariate-augmented specification, covariates contribute additional predictive information, and the marginal contributions of differential-quotient indicators are often near zero, while the early difference-quotient indicators remain among the most informative predictors (Table 8). For ease of comparison across model specifications, Table 9 summarizes the top predictors for at-risk students. The strongest overall predictors are the difference-quotient stages (Stages 1–3), followed by learner prerequisites (prior knowledge, domain-specific prior knowledge, and cognitive abilities), with only modest contributions from later stages.

7 Discussion

This study investigated how accurately, and at which earliest point along the theory-derived HLT, ML models can identify students at risk of acquiring pseudostructural knowledge of the derivative concept, using cumulative prediction stages that reflect the evidence available during instruction. Furthermore, we examined whether adding learner prerequisites (covariates) improves this identification and which ILT indicators contribute most to prediction.

RQ1: identification of students at-risk with ML and ILTS Throughout the discussion, prediction stages refer to cumulative points in the instructional sequence (Stages 1–9), whereas ILT indicators denote the phase-subconcept performance aggregates (e.g., DC × INT) used as predictors. ILT-based models achieved high recall throughout the instructional sequence, which is consistent with a screening-oriented use case in which missing as few at-risk learners as possible is prioritized over avoiding false alarms (Baker and Inventado 2014). However, this sensitivity came with modest discrimination and precision values (see Table 3). This suggests that ILT-based models can identify a large proportion of at-risk students early on, but only at the cost of a substantial number of false positives. Put differently, the models are useful for broad early screening, but not for precise individual diagnosis. One approach to address this was to evaluate model performance under a fixed precision floor of 0.30 (see Table 4). As expected, this more conservative operating rule reduced recall relative to the unconstrained F2-optimized setting, making the practical implications of operating-point choice explicit for this screening use case (e.g., Bowers et al. 2013).

Furthermore, these results align with Sfard (1991, 1992) process-object-approach. The difference quotient, particularly during its interiorization phase, represents the point at which learners begin to internalize operational knowledge of the difference quotient (e.g., Weber et al. 2012), which forms an essential foundation for subsequent structural knowledge (e.g., Litteck et al. 2025). Difficulties emerging at these phases therefore indicate early obstacles in the transition from operational to structural knowledge. Learners who fail to acquire knowledge about the concept of the difference quotient are unlikely to acquire operational or structural knowledge in later subconcepts. Instead, they may remain at the level of pseudostructural knowledge (Litteck et al. 2025; Sfard 1992).

Taken together, these findings suggest that the earliest practically informative point for identifying learners at risk of acquiring only pseudostructural knowledge is as early as the difference-quotient interiorization indicator (i.e., prediction Stage 1), and that later stages do not yield uniformly better identification. Under more restrictive precision-constrained decision rules, some later stages may yield selective advantages, but these improvements are not uniform across the instructional sequence and come with the disadvantege of being substantially later, allowing only later support for learners. The ILT-based indicators provide diagnostically valuable information overall, although precision remains limited under F2 optimization and improves only selectively under the precision-constrained operating rule (see Tables 3 and 4).

RQ2: added value of covariates to identify at-risk students Adding covariates yielded selective improvements in discrimination and changed the stage-specific pattern under the precision-constrained setting, consistent with the established role of learner prerequisites for mathematics learning and learning trajectories (Bednorz et al. 2025; Confrey et al. 2014; Hambrick and Engle 2002; Schneider and Preckel 2017). From a theory perspective, this is compatible with Sfard's (1991, 1992) view that concept development builds on available "concrete objects" that support interiorization and condensation processes. However, prerequisites alone did not fully account for predictive performance: compared with the prerequisites-only baseline (Stage 0), models incorporating ILT evidence showed only modest gains in global performance metrics, indicating that process information from the learning environment adds some incremental value beyond baseline prerequisite measures (see Table 5). In practical terms, this suggests that screening during instruction can make primary use of the available ILT indicators, while prerequisite measures may serve as a complementary source of information rather than a substitute for process-based evidence. Substantively, this result aligns with the idea that prerequisite knowledge (e.g., understanding slope of a secant line as grounding for the difference quotient and subsequent differential-quotient reasoning) becomes consequential insofar as it is expressed in learners' performance (Weber et al. 2012; Yu 2023). Under the precision-constrained setting, covariates did not yield a uniform advantage across later stages. Instead, the results suggest stage-specific shifts in where a more favorable precision-recall balance could be achieved. One possible interpretation is that prerequisite-related differences become especially relevant at specific stages. For instance, the selective gain at Stage 5 may reflect the increasing importance of prior knowledge when learners have to condense and coordinate the limit process underlying the differential quotient. Similarly, the gain at Stage 7 may be linked to the transition from a pointwise view of derivative-related reasoning to a functional view of the derivative function, a shift that may place greater demands on prior knowledge about functions and functional relationships. Although these interpretations remain speculative, they are consistent with the view that prerequisite variables can become particularly informative when learners are required to reorganize and extend their emerging conceptual structures (e.g., Litteck et al. 2023).

RQ3: which phase-subconcept ILT indicators are most important for predicting at-risk status? The feature importance analysis revealed two complementary dimensions that inform the identification of students at risk of acquiring pseudostructural knowledge: (a) what features are important for prediction, and (b) when within the unit these features become relevant.

In the ILT-only specification, the strongest predictors are the early difference-quotient ILT indicators, DC × INT and DC × CON (introduced at prediction Stages 1 and 2), followed by DC × REI (Stage 3) (see Tables 7 and 9). This permutation-importance pattern closely aligns with the logic of Sfard's hierarchical account (Sfard 1991). Learners first interiorize and condense processes related to the difference quotient, thereby establishing operational foundations from which structural knowledge can emerge. This, in turn, supports subsequent development in later subconcepts such as the differential quotient. Accordingly, while predictive information is distributed across predictors, it is particularly concentrated in the early difference-quotient indicators (cf., Weber et al. 2012).

In the covariate-augmented model, the structure of feature importance changed substantially. Covariates emerged as important predictors, underscoring the crucial role of learner prerequisites in explaining why students struggle with concept acquisition (Bednorz et al. 2025; Hambrick and Engle 2002; Schneider and Preckel 2017; Stern 2015) (see Tables 8 and 9). These variables account for interindividual differences that the ILT indicators alone could not explain. The results show that prior knowledge accounts for the largest covariate effect (e.g., Stern 2015). Notably, even with covariates included, the DC × CON ILT indicator retains the highest mean importance, underscoring the epistemic significance of concept acquisition within Sfard's framework and across the subconcepts of the concept of derivative (Sfard 1991; Zandieh 2000). Like the ILT-based model, the results align with Sfard's (1991) hierarchical view of concept formation. However, adding covariates appears to make some later indicators, particularly those associated with the derivative function (DF × INT, DF × CON, DF × REI), more relevant within the overall predictive pattern. The observed shift from ILT-based indicators to covariates is consistent with the idea that weaker prior knowledge constrains the execution of the necessary operations on "concrete objects" (Sfard 1991, p. 13), thereby impeding the acquisition of operational knowledge in the first place. In addition, with covariates most predictive information concentrates on the difference quotient and derivative function phases, while differential quotient phases contribute little marginal importance (see Table 8). This result suggests that the difference quotient may already capture much of the variance relevant for later outcomes, while additional variance is accounted for by prior knowledge and cognitive abilities, leaving the differential-quotient indicators with less incremental predictive information.

8 Limitation and conclusion

For this study, several limitations must be acknowledged. First, because the ILT indicators summarize correctness under instructional conditions (including feedback and support), they should be interpreted as performance evidence within the learning environment rather than as decontextualized measures of individual mastery. Relatedly, our ILT indicators rely on a deliberately parsimonious dichotomous coding (correct/incorrect) at the activity level. This operationalization does not exploit richer process evidence that digital learning environments can in principal provide (e.g., intermediate steps, response pathways, timing, or distractor-level information), which may contain additional diagnostic information for theory-aligned learning-process inferences and at-risk screening. Furthermore, models were tuned for high sensitivity under a fixed precision target, prioritizing recall over specificity. Accordingly, variations in precision across cumulative predictor sets reflect an intended trade-off rather than model deterioration. Third, limit-related items are underrepresented in the prerequisites measures, which may partly explain their weaker incremental contribution at points in the instructional sequence where limit understanding is central (e.g., the differential quotient). Fourth, the sample size was modest ($n = 108$) and performance was evaluated via internal cross-validation only. This restricts generalizability and may yield optimistic estimates. Precision-recall patterns are also conditioned on the at-risk base rate in the present sample and may shift under different prevalences. Finally, the focus on one mathematical topic (the derivative) and one specific digital learning environment constrains transfer to other domains and instructional designs. The robustness analyses further showed that continuous posttest performance was only weakly predictable from the available predictors, which may help explain why predictive gains remained moderate overall.

Taken together, the results illustrate a practical and theory-sensitive use of digital learning environments for tracing students' acquisition of knowledge about the derivative. From an applied perspective, high-recall models based on early learning data can serve as broad, low-stakes screening tools to minimize missed at-risk learners, but this comes at the cost of false positives. Consequently, ILT-based prediction should be used as an initial signal to trigger follow-up assessment and targeted support, rather than as a stand-alone decision rule. When available, learner prerequisites (e.g., prior knowledge and cognitive abilities) provide complementary information that can improve discrimination, yet such measures often require additional assessment time and may not be routinely available in school practice.

Importantly, our results primarily locate when along the theory-based instructional trajectory screening-relevant evidence emerges. Moreover, they do not directly test specific instructional mechanisms. Within this boundary, the concentration of predictive information in early difference-quotient indicators suggests that instruction should prioritize robust operational knowledge and the transition toward structural knowledge for the difference quotient, especially when instructional time is limited and teachers must make pragmatic choices about emphasis. Drawing on established mathematics education accounts, this includes supporting students in moving from a static, linear-only view of rate of change toward a more flexible and dynamic perspective that spans different function types and varying interval widths (Weber et al. 2012; Zandieh 2000). In the subsequent development toward the differential quotient, targeted scaffolding can explicitly support the process-object transition, particularly during condensation, where learners begin to coordinate representations and meanings more formally (Cottrill et al. 1996; Sfard 1991, 1992). Instructional approaches such as contrasting cases, dynamic visualizations, and activities that foreground limit-based reasoning may facilitate this shift and prepare the ground for later formalization (Cornu 1991; Cory and Smith 2011; Cottrill et al. 1996). Later phases should continue to promote reification by consolidating the object-like status of the difference quotient, differential quotient, and derivative function (Litteck et al. 2025; Sfard 1991; Zandieh 2000).

Finally, even in this theory-based and content-rich setting, discrimination and precision remained modest. This provides an empirically grounded boundary condition: early warning based on process indicators can be informative for sensitivity-oriented screening, but will likely not yield highly precise individual diagnosis.

Supplementary Information The online version of this article (https://doi.org/10.1007/s42010-026-00245-z) contains supplementary material, which is available to authorized users.

Funding This work was supported by the Leibniz Association under grant number K365/2020.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The dataset generated and analyzed during the presented study is available from the corresponding author upon reasonable request.

Declarations

Conflict of interest D. Bednorz, D. Sommerhoff and A. Heinze declare that they have no competing interests.

Ethical standards Ethical approval was obtained for the presented study from the Ethics Committee of the IPN—Leibniz Institute for Science and Mathematics Education under 2022_33_KU. Parental consent was obtained for all participants.

References

Alnasyan, B., Basheri, M., & Alassafi, M. (2024). The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *6*, 100231. https://doi.org/10.1016/j.caeai.2024.100231.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: a systematic literature review. *Computers & Education*, *141*, 103612. https://doi.org/10.1016/j.compedu.2019.103612.

Anghel, E., Khorramdel, L., & von Davier, M. (2024). The use of process data in large-scale assessments: a literature review. *Large-Scale Assessments in Education*. https://doi.org/10.1186/s40536-024-00202-1.

Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235.

Azizah, Z., Ohyama, T., Zhao, X., Ohkawa, Y., & Mitsuishi, T. (2024). Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data. *Computers and Education: Artificial Intelligence*, *7*, 100261. https://doi.org/10.1016/j.caeai.2024.100261.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: from research to practice* (pp. 61–75). Springer.

Barrett, J. E., Sarama, J., Clements, D. H., Cullen, C., McCool, J., Witkowski-Rumsey, C., & Klanderman, D. (2012). Evaluating and improving a learning trajectory for linear measurement in elementary

grades 2 and 3: a longitudinal study. *Mathematical Thinking and Learning*, *14*(1), 28–54. https://doi.org/10.1080/10986065.2012.625075.
Bednorz, D., Sommerhoff, D., & Heinze, A. (2025). Acquiring a concept and its sub-concepts: Insights into the relevance of prerequisites for learning. In C. Cornejo et al. (Eds.), Proceedings of the 48th Conference of the International Group for the Psychology of Mathematics Education: Research reports (pp. 107–114).
Berry, J. S., & Nyman, M. A. (2003). Promoting students' graphical understanding of the calculus. *The Journal of Mathematical Behavior*, *22*(4), 479–495. https://doi.org/10.1016/j.jmathb.2003.09.006.
Bjerre, L. M., Peixoto, C., Alkurd, R., Talarico, R., & Abielmona, R. (2024). Comparing AI/ML approaches and classical regression for predictive modeling using large population health databases: applications to COVID-19 case prediction. *Global Epidemiology*, *8*, 100168. https://doi.org/10.1016/j.gloepi.2024.100168.
Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, *96*(2), 77–100. https://doi.org/10.1353/hsj.2013.0000.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.
Brodie, K. (2014). Learning about learner errors in professional learning communities. *Educational Studies in Mathematics*, *85*(2), 221–239. https://doi.org/10.1007/s10649-013-9507-1.
van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v045.i03.
Chernikova, O., Sommerhoff, D., Stadler, M., Holzberger, D., Nickl, M., Seidel, T., Kasneci, E., Küchemann, S., Kuhn, J., Fischer, F., & Heitzmann, N. (2025). Personalization through adaptivity or adaptability? A meta-analysis on simulation-based learning in higher education. *Educational Research Review*, *46*, 100662. https://doi.org/10.1016/j.edurev.2024.100662.
Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, *6*(2), 81–89. https://doi.org/10.1207/s15327833mtl0602_1.
Clements, D. H., & Sarama, J. (2020). *Learning and teaching early math* (2nd edn.). Routledge. https://doi.org/10.4324/9781003083528.
Confrey, J., Maloney, A. P., & Corley, A. K. (2014). Learning trajectories: a framework for connecting standards with curriculum. *ZDM*, *46*(5), 719–733. https://doi.org/10.1007/s11858-014-0598-7.
Cornell-Farrow, S., & Garrard, R. (2020). Machine learning classifiers do not improve the prediction of academic risk: evidence from Australia. *Communications in Statistics: Case Studies, Data Analysis and Applications*, *6*(2), 228–246. https://doi.org/10.1080/23737484.2020.1752849.
Cornu, B. (1991). Limits. In D. Tall (Ed.), *Advanced mathematical thinking* (pp. 153–166).
Cory, B., & Smith, K. W. (2011). Delving into limits of sequences. *The Mathematics Teacher*, *105*(1), 48–55. https://doi.org/10.5951/mathteacher.105.1.0048.
Cottrill, J., Dubinsky, E., Nichols, D., Schwingendorf, K., Thomas, K., & Vidakovic, D. (1996). Understanding the limit concept: beginning with a coordinated process scheme. *The Journal of Mathematical Behavior*, *15*(2), 167–192. https://doi.org/10.1016/S0732-3123(96)90015-2.
Dong, J., Sun, R., Yan, Z., Shi, M., & Bi, X. (2025). Research on learning achievement classification based on machine learning. *PLOS ONE*, *20*(6), e325713. https://doi.org/10.1371/journal.pone.0325713.
Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*(6), 284–292. https://doi.org/10.3102/0013189X17725747.
Eegdeman, I., Cornelisz, I., Meeter, M., & van Klaveren, C. (2023). Identifying false positives when targeting students at risk of dropping out. *Education Economics*, *31*(3), 313–325. https://doi.org/10.1080/09645292.2022.2067131.
Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education*, *106*, 83–96. https://doi.org/10.1016/j.compedu.2016.12.001.
Faucon, L., Olsen, J. K., Haklev, S., & Dillenbourg, P. (2020). Real-time prediction of students' activity progress and completion rates. *Journal of Learning Analytics*, *7*(2), 18–44. https://doi.org/10.18608/jla.2020.72.2.
Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. https://jmlr.org/papers/v20/18-760.html.

Foody, G. M. (2023). Challenges in the real world use of classification accuracy metrics: from recall and precision to the Matthews correlation coefficient. *PLOS ONE*, *18*(9), e291908. https://doi.org/10.1371/journal.pone.0291908.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x.

Gobert, J. D., Sao Pedro, M. A., Raziuddin, J. J., & Baker, R. S. (2013). From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391.

Hakimi, L., Eynon, R., & Murphy, V. (2021). The ethics of using digital trace data in education: a thematic review of the research landscape. *Review of Educational Research*, *91*(5), 671–717. https://doi.org/10.3102/00346543211020116.

Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: an investigation of the knowledge-is-power hypothesis. *Cognitive Psychology*, *44*(4), 339–387. https://doi.org/10.1006/cogp.2001.0769.

He, S., & Cui, Y. (2025). A systematic review of the use of log-based process data in computer-based assessments. *Computers & Education*, *228*, 105245. https://doi.org/10.1016/j.compedu.2025.105245.

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: a context-specific meta-analysis. *Computers & Education*, *153*, 103897. https://doi.org/10.1016/j.compedu.2020.103897.

Hilz, A., Hofman, A., Jansen, B., & Aldrup, K. (2025). Tracing students' practice behavior in an adaptive math learning program: does it mediate the math anxiety-performance link? *Learning and Instruction*, *98*, 102113. https://doi.org/10.1016/j.learninstruc.2025.102113.

Hlosta, M., Zdráhal, Z., & Zendulka, J. (2017). Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. LAK '17. (pp. 6–15). Association for Computing Machinery. https://doi.org/10.1145/3027385.3027449.

Iwagami, M., Inokuchi, R., Kawakami, E., Yamada, T., Goto, A., Kuno, T., Hashimoto, Y., Michihata, N., Goto, T., Shinozaki, T., Sun, Y., Taniguchi, Y., Komiyama, J., Uda, K., Abe, T., & Tamiya, N. (2024). Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: a development and validation study. *PLOS Digital Health*, *3*(8), e578. https://doi.org/10.1371/journal.pdig.0000578.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education*, *104*, 18–33. https://doi.org/10.1016/j.compedu.2016.10.001.

Kleiner, I. (2001). History of the infinitely small and the infinitely large in calculus. *Educational Studies in Mathematics*, *48*(2–3), 137–174. https://doi.org/10.1023/A:1016090528065.

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Neumann, K., & Rummel, N. (2022). Toward learning progression analytics: Developing learning environments for the automated analysis of learning using evidence centered design. Frontiers in Education, 7, Article 981910. https://doi.org/10.3389/feduc.2022.981910

Kucirkova, N., Gerard, L., & Linn, M. C. (2021). Designing personalised instruction: a research and design framework. *British Journal of Educational Technology*, *52*(5), 1839–1861. https://doi.org/10.1111/bjet.13119.

Lam, P. X., Mai, P. Q. H., Nguyen, Q. H., Pham, T., Nguyen, T. H. H., & Nguyen, T. H. (2024). Enhancing educational evaluation through predictive student assessment modeling. *Computers and Education: Artificial Intelligence*, *6*, 100244. https://doi.org/10.1016/j.caeai.2024.100244.

Lavelle-Hill, R., Frenzel, A. C., Goetz, T., Lichtenfeld, S., Marsh, H. W., Pekrun, R., Sakaki, M., Smith, G., & Murayama, K. (2024). How the predictors of math achievement change over time: a longitudinal machine learning approach. *Journal of Educational Psychology*, *116*(8), 1383–1403. https://doi.org/10.1037/edu0000863.

Litteck, K., Rolfes, T., & Heinze, A. (2023). More than just the basic derivation formula: the impact of prior knowledge on the acquisition of knowledge about the concept of derivative. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel & M. Tabach (Eds.), *Proceedings of the 46th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 283–290). PME.

Litteck, K., Rolfes, T., & Heinze, A. (2025). The structure of knowledge about the concept of derivative—a study investigating a process-object framework. *International Journal of Mathematical Education in Science and Technology*, *56*(11), 2164–2184. https://doi.org/10.1080/0020739X.2024.2397990.

Lou, Y., & Colvin, K.F. (2025). Performance prediction using educational data mining techniques: a comparative study. *Discover Education*, *4*, 112. https://doi.org/10.1007/s44217-025-00502-w.

Lu, Y., Wang, D., Chen, P., Meng, Q., & Yu, S. (2023). Interpreting deep learning models for knowledge tracing. *International Journal of Artificial Intelligence in Education*, *33*(3), 519–542. https://doi.org/10.1007/s40593-022-00297-z.

Mattison, R.E., Woods, A.D., Morgan, P.L., Farkas, G., & Hillemeier, M.M. (2022). Longitudinal trajectories of reading and mathematics achievement for students with learning disabilities. *Journal of Learning Disabilities*, *56*(2), 132–144. https://doi.org/10.1177/00222194221085668.

Mislevy, R.J., & Haertel, G.D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x.

Papamitsiou, Z., & Economides, A.A. (2014). Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64. http://www.jstor.org/stable/jeductechsoci.17.4.49.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pham, N., Ngoc, H.P., & Nguyen-Duc, A. (2025). Fairness for machine learning software in education: a systematic mapping study. *Journal of Systems and Software*, *219*, 112244. https://doi.org/10.1016/j.jss.2024.112244.

Plass, J.L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943.

Powers, D.M.W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*. arXiv. https://doi.org/10.48550/arXiv.2010.16061.

Riestra-González, M., Paule-Ruíz, M.P., & Ortín, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, 104108. https://doi.org/10.1016/j.compedu.2020.104108.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: an updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368–384. https://doi.org/10.1016/j.compedu.2007.05.016.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e118432. https://doi.org/10.1371/journal.pone.0118432.

Sarama, J., & Clements, D.H. (2009). *Early childhood mathematics education research: learning trajectories for young children*. Routledge.

Sarama, J., Clements, D.H., Barrett, J., van Dine, D.W., & McDonel, J.S. (2011). Evaluation of a learning trajectory for length in the early years. *ZDM*, *43*(5), 667–680. https://doi.org/10.1007/s11858-011-0326-5.

Schiller, R., Fleckenstein, J., Mertens, U., Horbach, A., & Meyer, J. (2024). Understanding the effectiveness of automated feedback: using process data to uncover the role of behavioral engagement. *Computers & Education*, *223*, 105163. https://doi.org/10.1016/j.compedu.2024.105163.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: a systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098.

Sfard, A. (1991). On the dual nature of mathematical conceptions: reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics*, *22*(1), 1–36. https://doi.org/10.1007/BF00302715.

Sfard, A. (1992). Operational origins of mathematical notions and the quandary of reification: the case of functions. In E. Dubinsky & G. Harel (Eds.), *The concept of functions: aspects of epistemology and pedagogy* (pp. 59–84). Mathematical Association of America.

Sfard, A., & Linchevski, L. (1994). The gains and the pitfalls of reification: the case of algebra. *Educational Studies in Mathematics*, *26*(2–3), 191–228. https://doi.org/10.1007/BF01273663.

Simon, M.A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, *26*(2), 114–145. https://doi.org/10.2307/749205.

Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: an elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning*, *6*(2), 91–104. https://doi.org/10.1207/s15327833mtl0602_2.

Sønderlund, A. L., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: a systematic review. *British Journal of Educational Technology*, *50*(5), 2594–2618. https://doi.org/10.1111/bjet.12720.

Stern, E. (2015). Intelligence, prior knowledge, and learning. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* 2nd edn. (Vol. 12, pp. 323–328). Elsevier.

Szilágyi, J., Clements, D. H., & Sarama, J. (2013). Young children's understandings of length measurement: evaluating a learning trajectory. *Journal for Research in Mathematics Education*, *44*(3), 581–620. https://doi.org/10.5951/jresematheduc.44.3.0581.

Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction. *Educational Researcher*, *41*(5), 147–156. https://doi.org/10.3102/0013189X12442801.

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: a dynamic framework. *Educational Psychology Review*, *33*(3), 863–882. https://doi.org/10.1007/s10648-020-09570-w.

Wang, K. D., Cock, J. M., Käser, T., & Bumbacher, E. (2023). A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. *British Journal of Educational Technology*, *54*(1), 192–221. https://doi.org/10.1111/bjet.13289.

Weber, E., Tallman, M., Byerley, C., & Thompson, P. W. (2012). Understanding the derivative through the calculus triangle. *The Mathematics Teacher*, *106*(4), 274–278. https://doi.org/10.5951/mathteacher.106.4.0274.

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, *112*, 106457. https://doi.org/10.1016/j.chb.2020.106457.

Wu, J.-Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement. *Computers & Education*, *163*, 104066. https://doi.org/10.1016/j.compedu.2020.104066.

Yu, F. (2023). Promoting productive understandings of rate of change in calculus courses. *PRIMUS*, *33*(9), 965–980. https://doi.org/10.1080/10511970.2023.2214891.

Zandieh, M. (2000). A theoretical framework for analyzing student understanding of the concept of derivative. In Dubinsky, A. H. Schoenfeld & J. Kaput (Eds.), *Research in collegiate mathematics education. IV* (Vol. 8, pp. 103–127). American Mathematical Society. https://doi.org/10.1090/cbmath/008/06.