Formative assessment in mathematics from pre-service preparation to the first year of teaching: A longitudinal multiple-case study

Tereza Kottová[1,*], Darina Jirotková[1]

[1] Faculty of Education, Charles University, Magdalény Rettigové 4, 116 39 Prague, Czech Republic; tereza.kottova@pedf.cuni.cz

This longitudinal multiple-case study investigates the development of formative assessment among five Czech pre-service primary mathematics teachers across university preparation and their first year of teaching after graduation. Using repeated interviews, lesson plan analysis, and video observations, the study examines how novices' understanding and classroom enactment of a five-strategy formative assessment (FA) framework develop over time. Findings indicate that while initial teacher education successfully initiates a conceptual shift toward process-oriented assessment, classroom enactment remains highly uneven. Practices supporting discussion and peer learning were most visible, while structured routines for eliciting evidence strengthened over time but remained uneven; explicit work with learning goals, success criteria, and criteria-linked feedback remained weak across phases. Key facilitators included collegial and leadership support, observation of concrete practices, and structured reflection; major constraints included time pressure, grading demands, parental expectations, and the demands of interpreting pupils' mathematical thinking in real time.

Key words: formative assessment, pre-service elementary teachers, mathematics education, teacher preparation, teaching practice.

Received 3/2026
Revised 6/2026
Accepted 6/2026

1 Introduction

Formative assessment (FA) is widely recognized as one of the most powerful instructional catalysts for pupil achievement and self-regulated learning (Black & Wiliam, 1998; Hattie & Timperley, 2007). In mathematics education, the call to integrate FA is particularly urgent. Assessing mathematical proficiency requires teachers to shift from evaluating the binary correctness of final answers to diagnosing underlying cognitive schemas, procedural errors, and pupil reasoning in real-time (Suurtamm et al., 2016). Recognizing this profound educational impact, international frameworks and national educational policies—such as the Czech Republic's Strategy for the Education Policy up to 2030+, increasingly support the implementation of FA as a core professional requirement for schools.

However, despite this strong theoretical and policy-level support, a critical implementation gap persists in daily school practice, particularly in STEM subjects (Schildkamp et al., 2020; Yan et al., 2021). This gap is also relevant in the Czech educational context, where FA has been discussed in the educational literature and where national policy increasingly emphasises changes in the content, methods, and assessment of education (Ministerstvo školství, mládeže a tělovýchovy [MŠMT], 2020; Starý & Laufková, 2016). In mathematics, the enactment of FA is especially demanding because it requires teachers to interpret pupils' solution strategies, errors, and mathematical reasoning rather than merely evaluate the correctness of final answers (Hošpesová, 2018; Suurtamm et al., 2016). International research further shows that developing this capacity is challenging for pre-service teachers, who need to learn how to use evidence of pupils' mathematical thinking to adapt instruction while also navigating contextual and practical constraints of classroom teaching (Ayalon & Wilkie, 2021; Van Orman et al., 2025).

This phenomenon points to a need for further longitudinal research on how FA develops during the transition from teacher education to professional practice. Recent syntheses in mathematics education reveal that research on FA remains empirically fragmented and frequently fails to capture the granular, process-oriented mechanisms required to interpret pupils' mathematical reasoning in real-time (Maskos et al., 2025). Furthermore, while numerous cross-sectional studies provide static snapshots of teacher candidates' beliefs, the field lacks continuous, longitudinal evidence tracking how the conceptualization and actual enactment of FA evolve as teachers cross the threshold from a scaffolded university environment into their first year of autonomous teaching (Van Orman et al., 2025). A deeper understanding of the specific cognitive and systemic dynamics of this transition could inform how teacher education programs design interventions to support sustained, subject-specific pedagogical change.

The present study contributes to addressing this gap by employing a longitudinal multiple-case study design. This research contributes to the field by systematically following the same cohort of elementary mathematics teachers from the theoretical coursework, through their practicum, and into their first year of professional teaching. This allows for a granular, context-sensitive account of how mathematics assessment practices develop, stabilize, or degrade over time.

To fully unpack this transition, the study is structured around three vital, interrelated dimensions. First, it examines the evolution of teachers' theoretical understanding of FA (RQ1), which is necessary to determine whether the foundational philosophy is sustained or eroded. Second, it investigates the actual classroom enactment of specific FA strategies (RQ2), moving beyond self-reported beliefs to identify—through video analysis—exactly which instructional micro-routines in mathematics are successfully adopted and which remain consistently challenging. Finally, it explores the contextual supports and constraints (RQ3) that mediate this process. Investigating these three dimensions together is essential not only for understanding the local realities of primary schools but also for providing actionable evidence for the global STEM education community striving to bridge the knowing-doing gap in teacher preparation.

2 Theoretical background

2.1 Operationalizing FA and the five-strategy framework

FA is commonly conceptualised not as a particular instrument or isolated classroom technique, but as a process through which evidence about pupils' learning is elicited, interpreted, and used to inform subsequent instructional action (Black & Wiliam, 2009; Heritage, 2007). FA is broadly defined as an ongoing process in which evidence of pupil learning is elicited, interpreted, and used by teachers, learners, or peers to adapt instruction and make informed decisions about the next instructional steps (Black & Wiliam, 2009; Heritage, 2007).

A practically useful way of specifying this process is offered by the five-strategy framework of FA, summarised in Table 1 (Leahy et al., 2005; Wiliam & Leahy, 2015).

Table 1: Five formative assessment strategies

Strategy	Description
1	Clarifying, sharing, and understanding goals and criteria
2	Organizing discussions/activities/tasks to gather evidence of learning
3	Providing feedback that moves learning forward
4	Activating pupils as sources of learning for each other
5	Activating pupils as "owners" of their learning

In this article, these five strategies of FA serve as the main analytic framework for examining how key components of FA become visible in elementary mathematics lessons.

To address the first research question regarding the evolution of teachers' perspectives, it is essential to define what constitutes "understanding of FA" in the context of this study. Drawing on Hiebert and Carpenter (1992), who emphasize that understanding involves connecting new knowledge to existing cognitive networks, we define participants' understanding of FA not merely as the ability to recall definitions, but as their conceptual grasp of FA principles. Specifically, this encompasses their ability to articulate the purpose of assessment beyond summative grading, their capacity to theoretically differentiate between traditional and process-oriented assessment, and their ability to name and describe specific FA strategies they intend to use in their practice. Thus, in this study, participants' understanding of FA is treated as an interview-based construct concerning how participants conceptualise, explain, and intend to use FA, whereas their actual classroom enactment of FA is analysed separately under RQ2.

2.2 The didactic complexity of FA in mathematics: Shifting from product to process-oriented assessment

The cognitive demands of FA are significantly magnified when contextualized within the specific didactic requirements of elementary mathematics. In this discipline, evidence of learning is not merely visible in the binary correctness of an answer, but crucially within the strategies, representations, explanations, and procedural errors through which pupils make their mathematical thinking available for interpretation (Suurtamm et al., 2016). Over the last decade, discussions regarding assessment in mathematics education have increasingly argued that teachers must look beyond final products and attend more closely to pupils' underlying solution processes (Maskos et al., 2025).

Process-oriented assessment focuses on how learners approach mathematical problems, allowing feedback to target specific parts of the solution process and address misconceptions as they emerge sequentially (Herbert et al., 2022; Maskos et al., 2025). FA in this subject, therefore, depends heavily on the teacher's *diagnostic competence*—their ability to attend to such evidence, interpret what it suggests about the pupil's mathematical cognitive schema, and decide how to respond instructionally (Black & Wiliam, 2009;

Jacobs et al., 2010). Recent work on error analysis in mathematics demonstrates that process-oriented approaches can reveal stable patterns of faulty reasoning that remain entirely invisible if a teacher only evaluates final products (Herbert et al., 2022; Maskos et al., 2025).

Consequently, FA in mathematics is not merely a general "communication skill"; it involves rigorous didactic decision-making grounded in the interpretation of mathematical thinking. In the Czech context, Hošpesová's (2018) research corroborates this complexity, demonstrating that introducing formative elements—such as peer assessment in inquiry-oriented instruction—encounters specific, subject-bound difficulties for both teachers and pupils, requiring targeted support and the internalization of entirely new classroom routines. Furthermore, achieving constructive alignment is critical; if assessments prioritize reasoning, modeling, and proofs, pupils and teachers will orient their practice toward critical thinking, whereas assessments that only measure procedural recall will cause pupils to default to surface-level strategies (Maskos et al., 2025).

2.3 The transition to practice: The implementation gap and the need for longitudinal inquiry

While national policy frameworks, such as the Czech Republic's Strategy 2030+, mandate FA as a core professional competence (MŠMT, 2020, 2023), macro-level policies do not guarantee micro-level classroom enactment (Alarcón & Lawn, 2018). In reality, a persistent implementation gap exists (Schildkamp et al., 2020; Yan et al., 2021). Although teacher candidates successfully cultivate foundational assessment literacy through explicit instruction during pre-service preparation, translating these ideals into subject-specific practice remains a profound challenge (Van Orman et al., 2025). During the transition into their first year of teaching, novice teachers often struggle to cope with complex classroom realities, such as time constraints and school grading policies, which frequently leads to the selective or superficial enactment of FA (Schildkamp et al., 2020; Whalen et al., 2019).

Despite this critical barrier, recent syntheses reveal that research on FA in mathematics remains empirically fragmented, often failing to capture the granular mechanisms required for real-time diagnostics (Maskos et al., 2025). Furthermore, existing literature consists largely of cross-sectional snapshots of teacher beliefs, lacking continuous, longitudinal evidence on how an individual's conceptualisation of FA evolves—and either solidifies or degrades—during the volatile transition into autonomous teaching (Van Orman et al., 2025). The present study addresses this void by employing a longitudinal multiple-case design, systematically tracking the same cohort from theoretical coursework into their first year of professional teaching to isolate which specific FA components are successfully translated into practice.

2.4 Goals of pre-service teacher education and assessing enactment

The overarching goal of initial teacher education is not merely the acquisition of theoretical knowledge, but the development of adaptive expertise that allows future teachers to flexibly apply pedagogical principles in real classroom settings (Darling-Hammond & Bransford, 2005). Identifying whether this goal has been achieved requires shifting the analytical focus from what teachers report in surveys to what they actually do—their instructional enactment (Grossman et al., 2009). Enactment involves the translation of theoretical concepts into core teaching practices and observable routines. Consequently, evaluating the success of pre-service education in the context of FA necessitates an examination of both the successful execution of these strategies (RQ2) and the systemic or contextual factors (RQ3) that mediate this transfer from university seminars to the complex reality of an elementary school classroom.

3 Research methodology and design

The study is situated within the course *Child and Mathematics* at the Faculty of Education, Charles University, specifically focusing on Module B, which introduces a formative approach to teaching mathematics in elementary education. The module consists of 12 sessions of 135 minutes and links an explicit introduction to the five FA strategies with practical examples from mathematics classrooms.

While participants produce a range of pedagogical outputs during the course (e.g., scenarios aligned with FA strategies, ongoing reflections, and a final portfolio including an annotated lesson plan), this study treats the course primarily as the contextual baseline for professional learning. The primary data capturing the participants' evolving understanding and enactment across the longitudinal timeline are derived from repeated semi-structured interviews and video-based evidence of classroom practice.

3.1 Aim and research questions

The aim of the project is to map and describe how pre-service teachers and novice teachers use FA tools/practices in mathematics instruction, and how their understanding of FA and its enactment develop during the *Child and Mathematics* course (Module B) and subsequently during teaching practice placements, as well as in the first year after graduation.

The study was guided by the following main research question:

> How do pre-service teachers' and novice teachers' understanding and enactment of FA in mathematics lessons develop across the course, subsequent field placements, and the first year of teaching, and which factors do participants identify as supportive or constraining?

This main question was further specified through three sub-questions:

> RQ1: How does participants' understanding of FA in mathematics change over the course and teaching placements?
>
> RQ2: Which FA tools/practices do pre-service teachers and novice teachers use when implementing FA in their mathematics teaching during placements and after completing teacher education?
>
> RQ3: Which factors facilitate participants' implementation of FA, and which factors hinder it (e.g., course-related supports, placement conditions, school context, mentoring, time pressures, and curricular demands)?

3.2 Design and participants

The study was designed as a longitudinal qualitative multiple-case study (Creswell, 2012; Flick, 2009) examining the development of pre-service elementary teachers' understanding of FA and its enactment in mathematics instruction. Five female participants ($N = 5$), identified as R1–R5 throughout the article, were followed from the early phase of the *Child and Mathematics course* (Module B) through to early-career teaching. All pre-service teachers enrolled in the module in the summer semester 2023/2024 at the Faculty of Education, Charles University, were invited to participate; five agreed to take part. Module B provided the instructional context in which participants were introduced to the five FA strategies; however, the study was not designed to test the course's causal effects. Rather, it aimed to map and understand participants' development and enactment over time.

The first author served as both course instructor and researcher, which required careful handling of potential power dynamics. The second author contributed to the conceptual framing of the study, methodological consultation, interpretation of findings, and critical revision of the manuscript.

Participation was voluntary, could be withdrawn at any time without consequences, and did not change any course requirements, learning activities, or assessment procedures. Grading followed the standard course criteria and was conducted in the same way for participants and non-participants. Data were reported in anonymised form to protect participants' identities.

To address the author's dual role as instructor and researcher, several steps were taken to reduce the risk of confirmatory interpretations (e.g., "looking for progress"). Coding was guided by a written manual with explicit decision rules, and an audit trail (analytic memos and episode logs) was maintained to document how interpretations were reached. During analysis, the author systematically searched for counterexamples (e.g., lessons in later phases where Strategy 1 or 3 remained absent despite expectations) and used these cases to challenge emerging claims. Finally, a second coder independently analysed 33% of the videos and consensus discussions led to refinements of the coding manual, supporting transparency and consistency.

At the beginning of the study, four participants were in their fourth year of the programme (R1, R3, R4, R5) and did not have continuous responsibility for a class; R2 was in her fifth year of the programme while also working as a class teacher, with one year of teaching experience. Thus, at the beginning of the study, she was simultaneously a pre-service teacher in terms of her programme status and a novice teacher in terms of her professional experience. For the purposes of this study, a "novice teacher" is broadly defined as an educator in their first to third year of professional practice (Farrell, 2012). Therefore, despite her one year of prior experience, R2 is classified as a novice teacher, as she was still navigating the beginning stages of her career.

3.3 Data collection

Data were collected in three phases (P1–P3) over two years in order to capture both the development of participants' understanding of FA (RQ1) and its enactment in mathematics instruction (RQ2), including

facilitating and constraining factors (RQ3). An overview of the phases is provided in Table 2. Repeated data collection with the same participants across three phases enables the documentation of change over time (the longitudinal dimension) and the comparison of the development of both understanding and enactment within individual cases as well as across cases.

Table 2: Overview of data collection phases (P1–P3)

Characteristic	Phase 1 (P1)	Phase 2 (P2)	Phase 3 (P3)
Term	February 2024	November/December 2024	November/December 2025
Study/professional context	Early stage of the course	Internship during the final year of study	Early-career teaching

To provide context for the analysed video-recorded lessons, and acknowledging that the enactment of process-oriented FA is dependent on the specific mathematical content being taught, Table 3 summarises the grade level and mathematical topic for each lesson across the three phases.

Table 3: Context of the analysed video-recorded mathematics lessons (grade level and topic)

		R1	R2	R3	R4	R5
P1	Grade	3.	3.	3.	4.	5.
	Topic	Introduction to division with remainders; operator tasks	Quadrilaterals and their properties	"Stepping" (krokování) and translating steps into equations	"Biland coins" and the binary system	Large numbers in the universe
P2	Grade	3.	5.	3.	4.	5.
	Topic	Working with squared (grid) paper	Manipulative geometry	Written subtraction ("Grandpa Lesoň's animals")	Word problems	Finding the area of a triangle on a square grid
P3	Grade	5.	1.	1.	1.	1.
	Topic	Non-convex polygons	"Stepping" (krokování) – short route	Addition (digit 5)	Cube constructions	Addition with carrying over 10; estimation; coin problems

Note. Lesson topics are translated from Czech; selected local task labels (e.g., "krokování", "Biland coins") are retained for specificity.

To enhance the credibility of the findings, data from multiple sources were triangulated following the chronological sequence of the pedagogical process: (a) participants' written lesson plans, (b) pre-class interviews, (c) 45 minute video recordings of mathematics lessons, and (d) reflective post-lesson interviews (Flick, 2009; Hendl, 2016). Triangulating these sources allowed for a robust comparison between participants' intended learning goals and assessment strategies formulated before the lesson (lesson plans and pre-class interviews), their actual classroom enactment (video), and their subsequent rationale and reflection on pupils' learning (post-lesson interviews).

First, participants' written lesson plans served as a foundational data source, collected prior to the lessons, to capture their intended learning goals and planned assessment strategies before entering the classroom (Wiliam & Leahy, 2015).

Second, semi-structured pre-class interviews were conducted to capture participants' conceptualisations of FA and their perceived support and challenges in implementing it in elementary mathematics (RQ1, RQ3). Across all phases, these interviews followed a common core of questions focused on: (a) how participants would describe the concept of FA, (b) their personal stance towards FA, (c) what supports (or could support) them in using FA in mathematics teaching, and (d) what they perceived as the greatest challenges of implementing FA in elementary mathematics (Hendl, 2005; Kvale & Brinkmann, 2009).

Third, 45 minute video recordings of mathematics lessons made it possible to capture the actual enactment of FA in naturally occurring classroom settings (Derry, 2007; Powell et al., 2003) and to conduct repeated analytic work with the recordings (Goldman et al., 2007). Potential reactivity to video recording was mitigated through repeated data collection over time and consistent, unobtrusive recording procedures (Flick, 2009). The camera was placed at the back of the classroom to capture primarily the teacher, the board, and part of the class; audio was recorded using a recording device, most often a mobile phone.

Finally, the post-lesson interview took place 7–10 days after the lesson (60–90 minutes) and was conducted as a reflective conversation linked to specific moments from the recorded lesson (focusing on participants' decision-making and interpretations in those episodes). These semi-structured interviews provided space to capture how participants interpreted pupils' responses and their own instructional decision-making (Hendl, 2005; Kvale & Brinkmann, 2009).

3.4 Data analysis

The analysis was guided by the research questions and tailored to the nature of each data source. To enhance transparency, the links between the research questions and the data are summarised in Table 4.

Table 4: Alignment of research questions and data sources

Research question	Primary data	Additional data
RQ1 – Development of understanding of FA	Pre-class and post-lesson interviews (P1–P3)	Lesson plans (P1–P3)
RQ2 – Use of tools/practices FA in mathematics teaching	Video recordings (P1–P3)	Lesson plans (P1–P3), post-lesson interviews (P1–P3)
RQ3 – Supporting and constraining factors	Pre-class and post-lesson interviews (P1–P3)	

Note. P1–P3 denote the three phases of data collection (P1 = February 2024; P2 = November/December 2024; P3 = November/December 2025).

3.4.1 Interviews and post-lesson interviews (RQ1, RQ3; complementing RQ2)

Interviews from phases P1, P2, and P3 (including post-lesson interviews) were transcribed verbatim and analysed using thematic coding. Interviews followed the same core set of questions across phases; probing questions were added as needed for clarification. Transcripts were anonymised and, for the purposes of longitudinal analysis, organised into a working table (matrix) that enabled tracking responses over time and across participants.

Coding focused on three main areas: (a) participants' understanding of FA in mathematics instruction (RQ1), (b) perceived supports and barriers to implementation (RQ3), and (c) explanatory accounts of specific instructional decisions observed in the videos (an interpretive layer complementing RQ2). The coding process followed a hybrid approach: initial broad categories were established a priori based on the research aims and the theoretical framework, followed by an inductive, data-driven generation of specific sub-codes. These codes were continuously revised and clustered into explanatory themes.

A longitudinal reading of the data enabled comparisons of themes across phases and the identification of changes over time (e.g., what participants considered "formative" at different stages, what they felt they were able or unable to do, and why). Quotations reported in the findings were selected for their relevance to the research question and were always checked against the wider transcript context.

3.4.2 Video recordings (RQ2)

Video recordings from phases P1, P2, and P3 (i.e., three lessons per participant) were analysed using directed qualitative content analysis (Hsieh & Shannon, 2005), with initial codes derived from the five-strategy FA framework (Wiliam & Leahy, 2015). The five strategies were operationalised into 29 observable categories, which formed an observation protocol and coding scheme for the video analysis (illustrative examples are provided in Table 5; the complete coding scheme with all 29 categories is provided in the Appendix).

Table 5: Examples of observable categories (codes) for the five FA strategies

Formative assessment strategy	Examples of observable categories (codes)
1. Clarifying, sharing, and understanding goals and criteria	Working with learning goals; Checking pupils' understanding of the goal; Using success criteria
2. Organizing discussions/activities/tasks to gather evidence of learning	Time for pupil discussion; Wait time; Control/diagnostic questions
3. Providing feedback that moves learning forward	Feedback focus: task/process vs. person; Valuing effort vs. ability; Pupil opportunity to respond to feedback
4. Activating pupils as sources of learning for each other	Working in pairs/groups; Peer feedback
5. Activating pupils as "owners" of their learning	Working with mistakes; Self-assessment / reflection on learning; Multiple solution methods

This granular breakdown (comprising 29 specific sub-codes in total) is crucial for answering RQ2, as it allows the analysis to move beyond broad statements about the five strategies and pinpoint exactly which specific micro-practices (e.g., providing wait time, structuring peer feedback, or utilizing mistakes) novice teachers successfully integrate or abandon over time. Consequently, these categories serve as the analytical framework for structuring and reporting the classroom enactment findings in the subsequent Results section.

Each of the 29 categories was rated within each lesson using a three-point scale capturing the degree of explicitness of enactment: 0 = not observed, 1 = mostly implicit/marginal (the practice occurred once or subtly, without explicit naming or an evident role in the lesson structure), 2 = explicit/systematic (the practice was clearly named, used repeatedly, and/or structured pupils' work or was linked to learning goals/success criteria). For each rating, a brief episode note and time stamp were recorded.

The 0–2 scale was not used as a quantitative measure of teaching "quality" or as a FA "score". Rather, it served as an analytic aid for a structured description of how explicitly particular practices appeared within a lesson and to support comparisons across phases and cases. Interpretation therefore remained qualitative, grounded in episode descriptions from the video data and supported by triangulation across data sources (video recordings, interviews, and lesson plans). Some indicators (e.g., sustained work with success criteria or pupil portfolios) typically unfold across sequences of lessons rather than within a single 45 minute lesson. Therefore, a rating of 0 indicates that the practice was not observable in the recorded lesson, not necessarily that it was absent from the teacher's longer-term repertoire.

The unit of analysis was the entire lesson (one video recording), while coding was conducted at the level of instructional segments/episodes in which a given practice occurred. For each category, the occurrence in the lesson and a brief note stamp documenting how the practice was enacted were recorded (multiple categories could occur within the same segment where appropriate). Category occurrences were logged continuously in a coding spreadsheet. A second coder independently applied the same coding scheme and the 0–2 scale to 5 of the 15 video recordings (33%), purposively selected to cover different data collection phases and contrasting cases (maximum variation). Disagreements were resolved through consensus discussion and led to refinements of decision rules in the coding manual. For each participant, a case matrix (P1–P2–P3) was created summarising which practices appeared over time, strengthened/weakened, or did not occur.

3.4.3 Lesson plans (RQ1, RQ2)

Written lesson plans from phases P1, P2, and P3 were analysed using directed qualitative content analysis to identify the planned use of FA strategies and intended learning goals. Specifically, the analysis focused on identifying the presence or absence of planned activities corresponding to the five FA strategies. This analysis provided insight into participants' pedagogical intentions before entering the classroom, serving as a comparative baseline for the enacted practices observed in the video recordings and reflected upon during the interviews.

3.4.4 Integration and trustworthiness

Coded outputs were compiled into case files (R1–R5) and synthesised in within-case matrices (P1–P2–P3) to support within- and cross-case comparison (Creswell, 2012). Trustworthiness was supported through longitudinal data collection, triangulation of sources, double-coding of a subset of videos with consensus resolution, and analytic memos documenting coding decisions.

3.4.5 Ethical considerations

Informed consent was obtained from all participants and participating schools. Video recordings were stored on a secure, access-restricted repository; anonymised identifiers are used throughout the text and identifying information was removed.

4 Results

To address the breadth of the longitudinal data and ensure a clear analytical focus, the findings are organised into subsections corresponding to the three research questions. To provide a general picture of the results, an overarching longitudinal pattern emerged across the data sources: As participants progressed from the initial course (P1) through their practicum (P2) and into their first year of autonomous teaching (P3), their theoretical understanding of FA deepened from a narrow focus on "verbal feedback" toward a process-oriented view of continuous regulation. However, their classroom enactment remained uneven. While peer-learning practices were consistently strong and discussion-rich activities were visible, structured elicitation routines remained uneven, and the explicit use of learning goals and success

criteria—and aligning feedback to those criteria—remained systematically weak across all phases, heavily constrained by the cognitive load of novice teaching and school-level pressures.

4.1 Development of participants' understanding of FA in mathematics (RQ1)

To address RQ1, and in line with our theoretical definition of "understanding" (see Section 2.1), we examined how participants' conceptual grasp and ability to articulate the purpose of FA evolved across the longitudinal phases (P1–P3) through pre- and post-lesson interviews. Overall, definitions gradually shifted from relatively narrow, feedback-focused meanings towards a broader, process-oriented understanding of FA as ongoing work during learning that informs next instructional steps and supports pupils' reflection and responsibility. At the same time, traces of earlier meanings persisted in some cases, particularly the association of FA with descriptive, individualised verbal feedback.

In the first interviews (P1), several participants described FA primarily as individualised descriptive verbal assessment—a developmental alternative to marks. For instance, R4 contrasted it with grading as informationally weak: "not just 'here is this mark'..." (R4, P1). At the same time, some early traces of a process-oriented view were already present; R2 described FA as support that helps a pupil move forward by clarifying "where they are, where they are going, and what they need..." (R2, P1).

By the next interview round during their internship (P2), definitions more clearly framed FA as assessment during learning and as a repertoire of strategies that makes next steps visible. R5 explicitly contrasted FA with summative assessment as "ongoing" work that helps pupils see where they are and what they can do next, while also informing the teacher (R5, P2). Participants increasingly used language of actionable improvement and progress mapping (e.g., "what steps they can take so that next time it will be better", R4, P2), suggesting a shift from end-point comments to a process distributed across the lesson.

In the beginning-teacher phase (P3), a further broadening was visible in several cases. R2 described FA as "an overarching approach" guiding instructional decision-making and next steps (R2, P3), while others highlighted pupils' reflection and growing independence (R5, P3). Pupils' agency was articulated more explicitly: "the pupil must be active and I accompany them as a teacher" (R4, P3).

Taken together, the pattern suggests development from seeing FA primarily as "better feedback" towards a more integrated view of FA as a principled way of organising mathematics teaching—connecting ongoing evidence about pupils' understanding with next instructional steps and pupils' active role.

4.2 Enactment of FA in the observed mathematics lessons (RQ2)

Across the five cases and three time points (15 video-recorded lessons—5 participants ×3 phases; 45 minutes each, triangulated with corresponding lesson plans), an uneven profile of FA enactment emerged. Table 6 provides a descriptive summary of the coding patterns across the five FA strategies and phases. The matrix serves as an analytic heuristic to reveal the overarching trajectory of classroom practice: peer-learning practices (Strategy 4) were consistently observed, while Strategy 2 showed a more uneven pattern: discussion-rich activities were visible, but structured elicitation routines strengthened mainly towards P3, while explicit work with learning goals, success criteria, and feedback aligned to these anchors (Strategies 1 and 3) remained largely absent or implicit. Strategy 5 (activating pupils as "owners of their learning") appeared selectively and strengthened slightly over time.

Table 6: Distribution of explicitness ratings (0–2) across FA strategies and phases

Strategy	P1 (02/2024)	P2 (11/2024)	P3 (11/2025)
S1 Goals & success criteria (7 indicators)	35/0/0	30/2/3	24/10/1
S2 Eliciting evidence (discussion & tasks) (8 indicators)	25/5/10	19/13/8	13/14/13
S3 Feedback that moves learning forward (4 indicators)	18/2/0	16/4/0	15/4/1
S4 Pupils as learning resources for one another (4 indicators)	3/8/9	7/3/10	2/6/12
S5 Pupils as owners of their learning (6 indicators)	14/10/6	8/15/7	6/14/10
All strategies (29 indicators)	95/25/25	80/37/28	60/48/37

Note. Cells show counts of lesson-level ratings of indicators as 0/1/2 (in this order), aggregated across five 45 min lessons (R1–R5) per phase. Strategy-level totals per phase correspond to the number of indicators in that strategy ×5 lessons. The table serves as a descriptive heuristic to illustrate enactment patterns over time and should not be interpreted as a quantitative quality score.

Building on the overarching trends presented in Table 6, the following subsections provide a detailed, step-by-step qualitative analysis of how each of the five FA strategies was enacted in the observed mathematics lessons. By triangulating video observations with participants' lesson plans and interview reflections, the specific qualitative nuances, shifts, and persistent challenges characterizing the transition from pre-service preparation to novice teaching are illustrated for each individual strategy.

4.2.1 Strategy 1: Clarifying learning goals and success criteria

This strategy involves defining clear learning intentions and criteria for success so that pupils understand what they are learning and what quality looks like. The triangulation of lesson plans and video data revealed a sharp contrast between intended and enacted curriculum. In P1, although learning goals were formally written in the participants' lesson plans (often broadly formulated and written in the third person), none of the observed lessons used them as a visible structuring element in the classroom. They were not communicated to pupils or revisited during the lesson.

In P2, the case of R4 was a notable exception: the teacher co-constructed a learning goal with pupils based on the task, wrote it on the board, and added a teacher-formulated goal. Although the formulation remained broad rather than mathematically specific, the goal was visibly used to frame and organise the lesson. Such enactment, however, remained isolated in the dataset.

In P3, some participants attempted to name the learning goals explicitly at the start of the lesson (e.g., opening questions such as "What will today's lesson be for?" or statements such as "Today your task will be to discover something…"). However, these moments often remained initial declarations and were not revisited; they were therefore mostly coded as limited/implicit rather than as structuring elements. Stronger enactment was observed when the first-year teacher returned to the goals during closing reflection (e.g., checking what pupils had actually discovered and learned). Across all cases and phases, work with success criteria was practically absent. Consequently, success criteria were unavailable as a support for self-assessment or for feedback linked to the quality of pupils' work.

4.2.2 Strategy 2: Eliciting evidence of learning through tasks and discussion

This strategy focuses on engineering classroom activities that make pupils' mathematical thinking visible to the teacher. Across all phases, most lessons provided ample space for pupils' activity, task solving, and sharing strategies. Teachers frequently used tasks that enabled comparison of methods, justification, and checking solutions, giving pupils time to articulate their thinking in pairs or whole-class discussions.

At the same time, more systematic "micro-techniques" for eliciting evidence (e.g., deliberate wait time, planned routines for quickly checking whole-class understanding, or explicit work with typical misconceptions) appeared less consistently. By P3, a modest strengthening of structured elicitation was visible, particularly when teachers built subsequent questions or tasks on pupils' strategies or on issues that emerged in discussion.

Vignette 1 (Phase 3, R1, Grade 5): Non-convex polygons (geoboards)

This vignette was selected because it represents one of the most explicitly enacted elicitation cycles (Strategy 2) in the dataset, demonstrating how a teacher can use emerging ambiguity in pupils' work to shape the next instructional move.

R1 used geoboards to elicit whole-class evidence: pupils repeatedly modelled and displayed shapes, enabling the teacher to scan responses and prompt peer noticing. She posed a constraint-based task (construct a quadrilateral with exactly two pegs outside) and invited pupils to compare and name their solutions. The solution to this task is illustrated in Fig. 1.

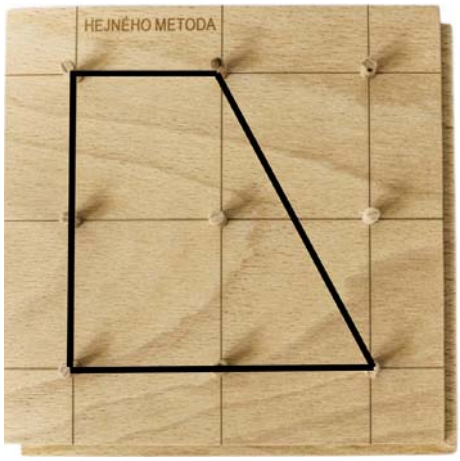

Fig. 1: Quadrilateral with exactly two pegs outside on a geoboard

An ambiguity emerged—whether pegs "touched" by the rubber band count as inside or outside—which R1 made public and clarified together with pupils. She then adapted the constraint, asking for a quadrilateral with exactly one peg outside. The pupils searched for solutions but were unsuccessful. Following a whole-class discussion on why such a construction is geometrically impossible under the given constraints, the teacher asked how the task could be modified to make a solution viable. The pupils suggested constructing a pentagon with exactly one peg outside; they subsequently discovered two distinct solutions and illustrated them on the blackboard (Fig. 2).

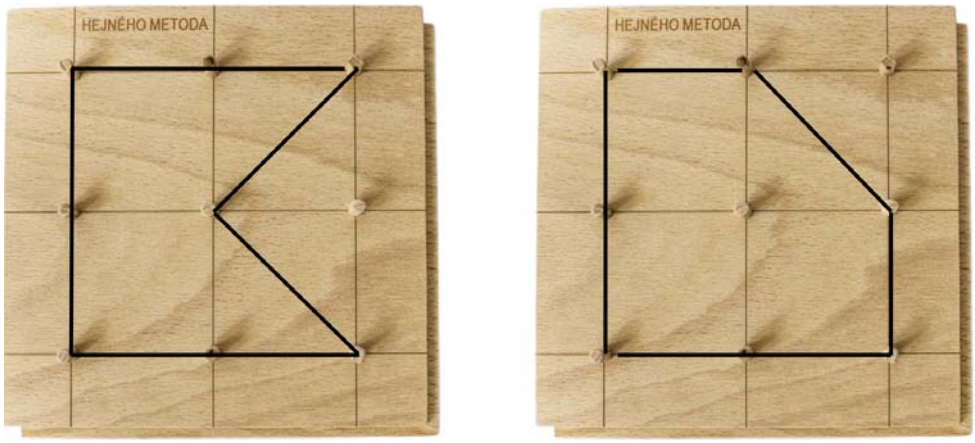

Fig. 2: Pupils' solutions for a pentagon with exactly one peg outside on a geoboard

She then used a pupil's informal label ("bitten-out") to bridge to more precise mathematical language and followed with classification prompts ("Which of these shapes does not belong?") to refine what does and does not count as non-convex. In the closing reflection, the class returned to the term non-convex polygon: pupils wrote (and several shared aloud) their own description of the concept in their own words. Overall, the vignette illustrates a level 2 (explicit/systematic) enactment: R1 repeatedly elicited whole-class evidence, sustained pupil-to-pupil comparison, and adjusted subsequent steps based on emerging ambiguities.

4.2.3 Strategy 3: Provision of feedback that moves learning forward

This strategy concerns providing feedback that focuses on the learning process and outlines concrete steps for improvement. The analysis of the coding matrix revealed that this was one of the least systematically enacted strategies. Feedback appeared predominantly as spontaneous comments, general praise, or organisational responses rather than as systematic feed-forward interventions.

In some lessons, teachers valued effort and process (e.g., work habits, collaboration, or approaches to solving), which supported a positive learning culture. However, in only two of the fifteen observed lessons did a teacher provide feedback that explicitly identified what was mathematically key in a pupil's solution, what counted as progress, and what a concrete next step should be. Crucially, feedback was not observed to be connected to shared learning goals or success criteria. Even when teachers summarised "what matters" in a task, this typically functioned as a general reminder rather than feedback anchored in explicitly shared criteria that pupils could use for self-regulation.

4.2.4 Strategy 4: Pupils as resources for one another

This strategy involves activating pupils as learning resources for their peers through collaboration and peer assessment. Activating pupils as learning resources for one another was the most frequently observed and most systematically enacted strategy across all participants (see Table 6). In all phases, the vast majority of lessons included pair or group work, sharing of solution methods, and situations in which pupils commented on classmates' strategies or corrected one another.

For example, in several P3 lessons, peer interaction functioned as a central organising principle of the lesson, as teachers deliberately built follow-up questions and whole-class discussion on pupils' solutions, explanations, and disagreements. However, in the absence of explicit success criteria (Strategy 1), peer feedback often remained at the level of simply comparing approaches ("I did it differently") rather than evaluating solutions against an agreed frame of "what counts as successful".

4.2.5 Strategy 5: Activation of pupils as owners of their learning

This strategy focuses on fostering self-regulation, reflection, and productive work with errors to make pupils active owners of their learning process. Elements of Strategy 5 appeared selectively across cases. Some lessons offered choices (e.g., method, level of challenge, mode of work), incorporated brief self-assessment, and treated errors as learning opportunities. In P3, these elements appeared more frequently for some participants and occasionally organised part of pupils' work (e.g., a structured end-of-lesson reflection). By contrast, pupil portfolios were observed in only one case (R2) as an additional tool rather than a stable routine.

Vignette 2 (Phase 2, R3, Grade 3): Written subtraction (52–48)

This vignette was selected because it illustrates a recurring pattern in the dataset: highly productive work with mistakes that surfaces mathematical misconceptions, but which ultimately lacks the necessary instructional follow-up to close the formative cycle.

R3 returned to a written subtraction problem (52–48) from the previous lesson after noticing recurring errors in pupils' work. She explicitly framed the revisiting as an opportunity to identify where and why mistakes occur. Pupils solved the problem on mini-whiteboards, discussed in pairs why it might be "tricky", and then shared their reasoning in whole-class talk. The discussion surfaced typical pitfalls (e.g., after borrowing, forgetting that the tens have decreased), and R3 made pupils' ideas public through questioning and confirmation. In this sense, the episode partially enacted a formative cycle: evidence from the previous lesson informed the selection of a follow-up task and the focus of the discussion. However, the work with mistakes remained limited: it did not include a subsequent check of understanding or a follow-up task (e.g., criteria-based checking or targeted practice) to consolidate the identified difficulty. The episode was therefore coded as a limited/implicit enactment of Strategy 5.

4.2.6 Cross-strategy summary: An uneven implementation profile

Overall, participants commonly planned and enacted discussion-rich lessons that supported pupils' activity and peer learning (Strategies 2 and 4) and included elements supporting self-regulation and productive work with errors (Strategy 5). At the same time, it remained highly challenging to anchor these practices in explicitly shared learning goals, success criteria, and feedback that builds on these anchors (Strategies 1 and 3). This pattern helps explain why lessons could appear "formative" through activity and discussion, yet only partly supported the systematic regulation of learning. Therefore, the interview data (RQ3) are examined to explore how participants interpreted these instructional choices and what contextual factors mediated their implementation.

4.3 Factors facilitating and constraining participants' implementation of FA in elementary mathematics (RQ3)

To address RQ3, participants' accounts of what facilitated and what hindered their use of FA in elementary mathematics across the longitudinal phases were analysed. The qualitative thematic analysis of interviews allowed for the identification and qualitative comparison of the specific contextual and cognitive barriers participants faced as they transitioned from pre-service to novice teachers.

4.3.1 Facilitating factors

The thematic analysis of the interview data facilitated the mapping of factors that participants consistently perceived as supportive across the three phases. The most prominent facilitating factor was a legitimising school context—particularly leadership and a collegial culture that supports pedagogical experimentation. Already in P1, participants anticipated the importance of a school that "functions as one organism" (R1, P1) and, as beginning teachers, described this more concretely as a climate "set up to assess and work formatively" (R2, P3).

A recurring facilitator was collegial support and sharing of practice, especially when colleagues were "excited to try something similar" and offered "tips and tricks" (R1, P2). Sharing was also framed as a mechanism of adoption: "Sharing with others helps me a lot—especially when I see it in action . . . if a colleague said 'we do this and it works, try it', I'd try it" (R5, P3).

Participants repeatedly highlighted the value of seeing FA enacted and having opportunities for rehearsal and joint analysis. Early on, one participant wished "to see it more" and learn how to translate theory into lesson preparation and questioning (R4, P1). Later, this became a call for repeated training

(including video-based work with teaching episodes) and protected space to try practices in one's own classroom (R2, P2).

The course experience supported implementation through concrete tools and a usable repertoire. For example, R3 described a checklist as "close to me" and "safe", because it provides clear criteria for both pupils and the teacher (R3, P2). Others reported immediate trials of specific techniques after seminars. Finally, participants stressed gradual implementation and ongoing learning. They described the need for a step-by-step pathway ("not to overdo it at the start") and, as beginning teachers, reported returning to their university course materials during planning: "Repetition is the mother of wisdom . . . when I plan, I have to look back to remind myself and be able to include it in teaching" (R2, P3).

4.3.2 Constraining factors

Thematic analysis mapped how cognitive and contextual constraints shifted during the first year of teaching. The most persistent barrier was time and capacity. Novices also struggled with diagnostic uncertainty, fear of change, and systemic pressures—particularly parental expectations for traditional grades—tersely summarised by one teacher as "Parents. School. System." (R4, P3).

Crucially, several constraints were mathematics-specific. Participants highlighted the intense cognitive load of diagnosing mathematical errors in real-time and orchestrating productive, non-binary classroom discussions. Furthermore, establishing explicit goals and linking them to self-assessment proved highly demanding, with teachers admitting to keeping goals "out of pupils' hands" due to their own uncertainty (R1, P3). Finally, routines like peer assessment were perceived as particularly fragile in early grades.

Taken together, these accounts explain the uneven enactment observed in the video data: while supports facilitated discussion and peer learning (Strategies 2 and 4), severe constraints around time, diagnostic uncertainty, and math-specific orchestration hindered the highly demanding work of setting explicit goals, success criteria, and providing criteria-linked feedback (Strategies 1 and 3).

4.4 Summary of findings for the main research question

Across phases (P1–P3), the five participants showed uneven but generally parallel development in their understanding of FA and its enactment in elementary mathematics. In interviews, their understanding shifted from a narrower view of "feedback instead of grades" (P1) towards a process-oriented conception linking evidence of learning to decisions about next steps and placing greater emphasis on pupils' active role (especially P3). The video analysis of 15 lessons similarly suggests a gradual strengthening over time, reflected in fewer "not observed" ratings and more implicit/explicit enactments across indicators. Practices supporting peer learning (Strategy 4) were consistently strongest. Discussion-rich activities (Strategy 2) were visible, whereas structured elicitation routines remained uneven and strengthened mainly over time. Self-regulation-oriented elements (Strategy 5) also strengthened selectively. By contrast, explicit work with learning goals and success criteria (Strategy 1; success criteria largely absent) and feedback aligned with these anchors (Strategy 3) remained weak and only rarely structured lessons. Participants identified school leadership and collegial support, opportunities to observe and rehearse concrete tools, and structured reflection on practice (e.g., video-based discussion) as key facilitators. The most frequently reported constraints were time pressure, grading and parental expectations, uncertainty in diagnosing learning and orchestrating mathematical discussion, and persistent difficulty anchoring instruction in explicit goals and success criteria.

5 Discussion

This longitudinal multiple-case study followed five participants across the transition from pre-service preparation into early-career teaching to examine the development of their understanding and enactment of FA in elementary mathematics. Drawing on repeated interviews, lesson plan analysis, and longitudinal video observations, the study documented a conceptual broadening alongside an uneven uptake of instructional practices. The significance of these findings lies not simply in cataloging what novice teachers do, but in illuminating the specific pedagogical and contextual fault lines where theoretical assessment literacy struggles to translate into sustainable classroom enactment.

5.1 The evolution of understanding (RQ1): The resilience of the grading paradigm

With regard to RQ1, participants' understanding of FA gradually shifted from a narrow, feedback-centered view (often equated with "verbal assessment instead of grades") toward a more robust, process-oriented

conception linking evidence of learning to subsequent instructional steps. This shift is in line with conceptualisations of FA that emphasise the use of evidence to adapt teaching and support further learning (Black & Wiliam, 1998, 2009). This development reflects a critical transition from perceiving assessment merely as an alternative form of evaluation to understanding its pedagogical function—how information is used to support mathematical progress.

Viewed through the lens of Hiebert and Carpenter (1992), who conceptualize understanding as the active construction of internal cognitive networks, this shift indicates that participants were not merely adopting new academic terminology. Rather, they were restructuring their pedagogical schemas, gradually connecting the isolated concept of FA to their broader networks of instructional decision-making and pupil agency.

This finding is important because it suggests that initial teacher education can contribute to a paradigm shift in how candidates conceptualize assessment. At the same time, this conceptual change proved to be gradual and occasionally unstable, with traces of early meanings—especially the strong association with individualized, descriptive feedback—persisting across phases. This instability is highly significant for teacher educators, as it points to the resilience of traditional grading norms. Assessment practices appear to be shaped by entrenched societal expectations and the enduring motivational consequences of traditional evaluation (Crooks, 1988; Koenka et al., 2021; Wisniewski et al., 2020). Therefore, the persistence of these early conceptions suggests that university programs may benefit from continuously addressing candidates' deeply socialized beliefs about the primary purpose of assessment. Furthermore, it highlights the broader challenge of transforming years of traditional schooling experiences within the scope of a single educational module.

5.2 Uneven enactment-strong discussion and peer learning, weak coherence of intentions-criteria-feedback (RQ2)

The video analysis showed an uneven enactment profile: peer learning was relatively strong, while discussion-rich practices were visible but structured elicitation routines remained uneven. Explicit work with learning intentions and especially success criteria remained weak, and feedback explicitly aligned with these anchors was limited. Interpreted in line with the five-strategy framework, this unevenness matters because learning intentions and success criteria provide a reference frame for interpreting evidence and making feedback usable for improvement (Leahy et al., 2005; Wiliam & Leahy, 2015).

The findings also indicate where coherence tends to break down for novices. Even when lessons include rich mathematical activity and pupil dialogue, without shared intentions and criteria it becomes difficult to translate elicited evidence and peer comments into concrete guidance about what to improve and how (Hattie & Timperley, 2007). In this sense, "active" classrooms may still lack anchors that make learning transparent and support systematic next steps for both pupils and teachers (Black & Wiliam, 2009; Heritage, 2007). This uneven implementation can lead to what has been described as pseudo-formative assessment (Jönsson, 2020), where the outward forms of active learning are present, but the essential structural and evaluative anchors are missing. Without explicitly shared criteria, lessons lack the necessary instructional coherence (Schmidt et al., 2005), making it highly difficult to consistently align classroom activities with intended mathematical learning outcomes.

5.3 Supports and constraints-conditions for sustaining formative routines (RQ3)

Findings related to RQ3 highlight that implementing FA is not merely an individual skill but a situated practice shaped by contextual "permission" and capacity. Participants repeatedly identified leadership support, collegial culture, opportunities to observe concrete enactments, and structured reflection (including video-based reflection) as key facilitators. Similar enabling conditions have been highlighted in the literature on teacher professional learning, particularly the role of collaborative inquiry and opportunities for systematic reflection on practice (Husu et al., 2008; OECD, 2005).

By contrast, persistent constraints included time pressure, grading demands and parental expectations, and uncertainty in diagnosing learning and orchestrating mathematical discussion. These barriers align with evidence that sustained enactment of FA requires support at both school and system levels and that translating understanding into stable routines is particularly challenging in the first years of teaching (Ayalon & Wilkie, 2021; OECD, 2005; Starý & Laufková, 2016). As the data indicate, the cognitive overload and systemic performance pressures of their new school environments hinder novices from translating their theoretical understanding of explicit criteria and feedback into stable classroom routines.

5.4 Limitations of the study

Several limitations should be acknowledged. First, the small, voluntary sample of five participants from a single programme provides in-depth, case-based insights rather than statistically representative claims, potentially over-representing individuals with a positive orientation toward FA (Creswell, 2012; Yan et al., 2021). Second, while the longitudinal design documents developmental trajectories, it cannot isolate the causal impact of the university course from other concurrent school experiences. Third, the reliance on a limited number of video recordings may miss periodic practices and remains subject to potential camera reactivity (Derry, 2007). Fourth, employing the five-strategy framework as a deductive coding lens—while ensuring comparability—necessarily backgrounds fine-grained diagnostic reasoning within mathematical content. Finally, lacking data on pupils' achievement, conclusions are limited to teachers' pedagogical development rather than measured impacts on student learning (Black & Wiliam, 1998; OECD, 2005).

5.5 Implications for teacher education, induction, and research

Overall, the findings suggest a partial closing though not elimination of the knowing-doing gap in FA in mathematics: participants' understanding broadened and their repertoire expanded, yet coherence-building components (especially success criteria and feedback anchored in intentions/criteria) remained consistently demanding. This pattern is consistent with frameworks that emphasise the interdependence of learning intentions, criteria, evidence, and feedback within formative assessment (Hattie & Timperley, 2007; Leahy et al., 2005; Wiliam & Leahy, 2015). Teacher education may therefore benefit from a stronger focus on these "hard-to-enact" mechanisms: translating lesson goals into pupil-accessible learning intentions, developing and using success criteria (not only goals), and building routines in which feedback and self-/peer assessment are explicitly anchored in intentions/criteria (Black & Wiliam, 2009; Heritage, 2007; Popham, 2008).

A second implication concerns the pedagogy of teacher learning. Repeated rehearsal, observation of concrete examples, and guided reflection (including video-based work with teaching episodes) can support novices in making decision-making visible and in stabilising routines that connect evidence, interpretation, and next steps (Hattie et al., 2017; Husu et al., 2008). In contexts where grading pressures are strong, induction and school-level support are particularly important for sustaining formative purposes alongside institutional expectations (MŠMT, 2020, 2023; Starý & Laufková, 2016). Understanding how this knowing-doing gap is negotiated during the induction phase may be relevant beyond the local context, particularly for STEM teacher education concerned with translating assessment theory into sustainable classroom practice.

For future research, it would be valuable to broaden the evidence base by including more diverse contexts and programmes, sampling teaching more frequently to capture longer-term routines, incorporating artefacts (pupils' work, tools for criteria), and–where feasible–linking analyses sensitively to indicators of pupils' learning (Black & Wiliam, 1998; OECD, 2005).

6 Conclusion

This longitudinal multiple-case study addresses an important gap in mathematics education research by tracking the fragile transition of pre-service teachers into their first year of autonomous practice. By mapping both the conceptual evolution and the classroom enactment of FA, the study moves beyond static snapshots to offer insights into the dynamic realities of learning to teach. The findings suggest that while initial teacher education can help initiate a conceptual shift—guiding novices away from traditional grading paradigms toward process-oriented assessment—the translation of these beliefs into stable routines remains uneven. Novices appear to readily adopt practices that generate active mathematical discourse and peer collaboration. However, the foundational anchors of FA—explicit success criteria, shared learning goals, and criteria-referenced feedback—are frequently challenged by the cognitive load of real-time mathematical diagnostics and the performance-oriented pressures of the school environment. Consequently, novice classrooms may sometimes lack the instructional coherence necessary to fully support self-regulated learning.

To help bridge this persistent knowing-doing gap, teacher education might benefit from certain structural adjustments. University programs could increasingly complement the cultivation of theoretical assessment literacy with repeated rehearsal of "hard-to-enact" micro-routines, such as the backward design of mathematical success criteria. Furthermore, the findings highlight that induction into the profession should ideally not be experienced as a period of isolated survival. Novice teachers would likely benefit from sustained, subject-specific mentoring and a supportive school culture that buffers their pedagogical

experimentation against entrenched grading norms. Ultimately, this study illustrates that the implementation of FA in mathematics is likely not merely an individual competency, but a deeply situated, systemic process. By identifying pedagogical and contextual conditions under which formative intentions may weaken during the transition into practice, this study contributes to broader discussions in STEM teacher education about how assessment can be used not only to measure learning, but also to support it.

References

Alarcón, C., & Lawn, M. (Eds.). (2018). *Assessment cultures: Historical perspectives*. Peter Lang.

Ayalon, M., & Wilkie, K. J. (2021). Investigating peer-assessment strategies for mathematics pre-service teacher learning on formative assessment. *Journal of Mathematics Teacher Education*, *24*(4), 399–426. https://doi.org/10.1007/s10857-020-09465-1

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31. https://doi.org/10.1007/s11092-008-9068-5

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4), 438–481. https://doi.org/10.2307/1170281

Darling-Hammond, L., & Bransford, J. (Eds.). (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do*. Jossey-Bass.

Derry, S. J. (2007). *Guidelines for video research in education: Recommendations from an expert panel*. Data Research and Development Center.

Farrell, T. S. C. (2012). Novice-service language teacher development: Bridging the gap between preservice and in-service education and development. *TESOL Quarterly*, *46*(3), 435–449. https://doi.org/10.1002/tesq.36

Flick, U. (2009). *An introduction to qualitative research* (4th ed.). SAGE.

Goldman, R., Pea, R., Barron, B., & Derry, S. J. (Eds.). (2007). *Video research in the learning sciences*. Routledge.

Grossman, P., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching*, *15*(2), 273–289. https://doi.org/10.1080/13540600902875340

Hattie, J., Fisher, D., Frey, N., Gojak, L. M., Delano Moore, S., & Mellman, W. (2017). *Visible learning for mathematics, grades K–12: What works best to optimize student learning*. Corwin.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hendl, J. (2005). *Kvalitativní výzkum: základní metody a aplikace* [Qualitative research: Basic methods and applications]. Portál.

Hendl, J. (2016). *Kvalitativní výzkum: základní teorie, metody a aplikace* [Qualitative research: Basic theory, methods, and applications]. Portál.

Herbert, S., Vale, C., White, P., & Bragg, L. A. (2022). Engagement with a formative assessment rubric: A case of mathematical reasoning. *International Journal of Educational Research*, *111*, Article 101899. https://doi.org/10.1016/j.ijer.2021.101899

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, *89*(2), 140–145. https://doi.org/10.1177/003172170708900210

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National council of teachers of Mathematics* (pp. 65–97). Macmillan.

Hošpesová, A. (2018). Formative assessment in inquiry-based elementary mathematics. In G. Kaiser, H. Forgasz, M. Graven, A. Kuzniak, E. Simmt, & B. Xu (Eds.), *Invited Lectures from the 13th International congress on mathematical education. ICME-13 Monographs* (pp. 249–268). Springer. https://doi.org/10.1007/978-3-319-72170-5_15

Husu, J., Toom, A., & Patrikainen, S. (2008). Guided reflection as a means to demonstrate and develop student teachers' reflective competencies. *Reflective Practice*, *9*(1), 37–51. https://doi.org/10.1080/14623940701816642

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Jacobs, V. R., Lamb, L. L. C., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, *41*(2), 169–202. https://doi.org/10.5951/jresematheduc.41.2.0169

Jönsson, A. (2020). Definitions of formative assessment need to make a distinction between a psychometric understanding of assessment and "evaluative judgment." *Frontiers in Education*, *5*, Article 2. https://doi.org/10.3389/feduc.2020.00002

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2021). A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology*, *41*(7), 922–947. https://doi.org/10.1080/01443410.2019.1659939

Kvale, S., & Brinkmann, S. (2009). *InterViews: Learning the craft of qualitative research interviewing* (2nd ed.). Sage Publications.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, *63*(3), 18–24. https://www.ascd.org/el/articles/classroom-assessment-minute-by-minute-day-by-day

Maskos, K., Schulz, A., Oeksuez, S. S., & Rakoczy, K. (2025). Formative assessment in mathematics education: A systematic review. *ZDM–Mathematics Education*, *57*(4), 679–693. https://doi.org/10.1007/s11858-025-01696-x

Ministerstvo školství, mládeže a tělovýchovy. (2020). *Strategie vzdělávací politiky ČR do roku 2030+* [Strategy for the education policy of the Czech Republic up to 2030+]. https://www.msmt.cz/vzdelavani/skolstvi-v-cr/strategie-2030

Ministerstvo školství, mládeže a tělovýchovy. (2023). *Kompetenční rámec absolventa a absolventky učitelství* [Competence framework for teacher education graduates]. https://www.msmt.cz/uploads/kompetencni_ramec_absolventa_2023_10.pdf

OECD. (2005). *Formative assessment: Improving learning in secondary classrooms.* OECD. https://www.oecd.org/education/ceri/35661078.pdf

Popham, W. J. (2008). *Classroom assessment: What teachers need to know.* Prentice Hall.

Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *The Journal of Mathematical Behavior*, *22*(4), 405–435. https://doi.org/10.1016/j.jmathb.2003.09.002

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, Article 101602. https://doi.org/10.1016/j.ijer.2020.101602

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, *37*(5), 525–559. https://doi.org/10.1080/0022027042000294775

Starý, K., & Laufková, V. (2016). *Formativní hodnocení ve výuce* [Formative assessment in teaching]. Portál.

Suurtamm, C., Thompson, D. R., Kim, R. Y., Diaz Moreno, L., Sayac, N., Schukajlow, S., Silver, E., Ufer, S., & Vos, P. (2016). *Assessment in mathematics education: Large-scale assessment and classroom assessment.* Springer. https://doi.org/10.1007/978-3-319-32394-7

Van Orman, D. S. J., Gotch, C. M., & Carbonneau, K. J. (2025). Preparing teacher candidates to assess for learning: A systematic review. *Review of Educational Research*, *95*(3), 427–463. https://doi.org/10.3102/00346543241233015

Whalen, C., Majocha, E., & Van Nuland, S. (2019). Novice teacher challenges and promoting novice teacher retention in Canada. *European Journal of Teacher Education*, *42*(5), 591–607. https://doi.org/10.1080/02619768.2019.1652906

Wiliam, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for K–12 classrooms.* Learning Sciences International.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042

Appendix Full coding scheme for video analysis

Operationalisation of the Five FA Strategies Into 29 Observable Categories (Codes)

Formative Assessment Strategy	Code	Category (observed indicator/question)
1. Clarifying, sharing, and understanding goals and criteria	1	Working with learning goals (How do the participants use lesson goals during the lesson?)
	2	Communicating the goal to pupils (How is the lesson goal communicated to pupils?)
	3	Formulating goals (How are goals formulated?)
	4	Checking pupils' understanding of the goal (How do the participants verify that pupils understand the goal?)
	5	Co-construction of goals (Are lesson goals co-constructed with pupils, or set only by the participants?)
	6	Working with success criteria (Do success criteria appear?)
	7	Using success criteria (How are criteria used, e.g., checklists, rubrics, examples, etc.?)
2. Organizing discussions/activities/tasks to gather evidence of learning	8	Time for pupil discussion (How much time/space is provided for pupils to discuss?)
	9	Discussion of solutions (Do pupils compare solutions, argue, and explain reasoning?)
	10	Selecting pupils to answer (How do the participants select who answers?)
	11	Responding to "I don't know" (How do the participants work with this response?)
	12	Time to think (wait time) (Is wait time provided; how long/when?)
	13	Using statements/paraphrases vs. questions (Do the participants use statements/paraphrases, rather than only asking questions?)
	14	Whole-class response system (Is there a system to quickly capture responses from the whole class, e.g., mini-whiteboards, traffic lights, voting?)
	15	Control/diagnostic questions (Are diagnostic questions used to check understanding? If yes, how?)
3. Providing feedback that moves learning forward	16	Feedback focus: task/process vs. person (Is feedback directed to the task/process or to the person?)
	17	Valuing effort vs. ability (Does feedback emphasise effort/strategy over fixed ability?)
	18	Pupil opportunity to respond to feedback (through revision/correction/follow-up action)
	19	Feedback linked to objectives/criteria (Is feedback explicitly linked to the lesson objectives and/or success criteria?)
4. Activating pupils as sources of learning for each other	20	Face-to-face interaction (Do pupils have opportunities for face-to-face interaction?)
	21	Working in pairs/groups (Are pair/group activities used?)
	22	Group formation (How do the participants form pairs/groups?)
	23	Peer feedback (Does peer feedback occur? If yes, how is it structured?)
5. Activating pupils as "owners" of their learning	24	Working with mistakes (How are mistakes handled as learning opportunities?)
	25	Pupils' reflection on mistakes (Do pupils have an opportunity to think about where/why mistakes occurred?)
	26	Self-assessment / reflection on learning (Is self-assessment/reflection present? If yes, how?)
	27	Opportunities for choice during the lesson (Do pupils have options, e.g., difficulty, pace, mode of work, tools, etc.?)
	28	Multiple solution methods (Are different ways of solving tasks enabled/supported?)
	29	Pupil portfolio (Is a pupil portfolio used? If yes, how?)