Scientia in educatione, 17(1), 2026
http://www.scied.cz
ISSN 1804-7106

Obsah

Výzkumné stati

Nakladatelství Karolinum

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.

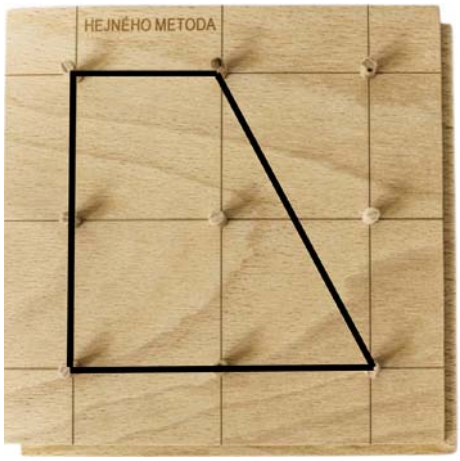

Fig. 1: Quadrilateral with exactly two pegs outside on a geoboard

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).

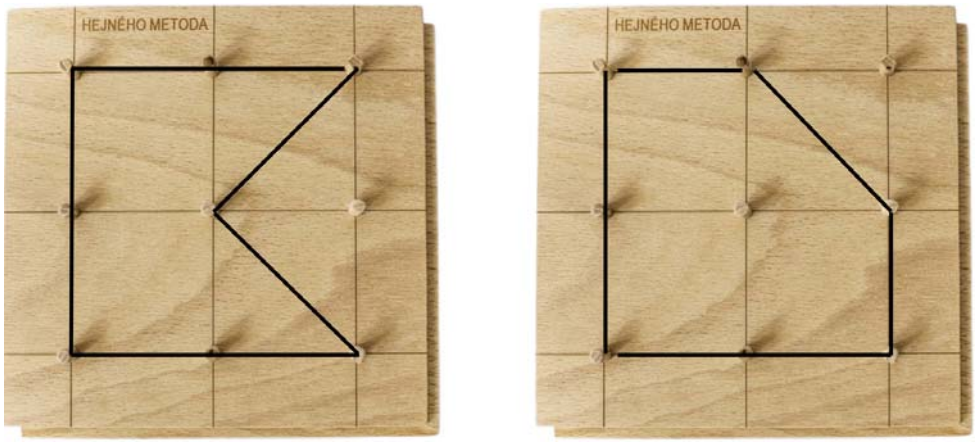

Fig. 2: Pupils' solutions for a pentagon with exactly one peg outside on a geoboard

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?)

Astronomical misconceptions among South African science teachers: A case study

Khutso Charles Mogale[1], Abraham Motlhabane[1,*]

[1] Department of Science and Technology Education, University of South Africa, Preller Street, Pretoria, South Africa; motlhat@unisa.ac.za

This study explores the astronomical misconceptions of 10 South African science teachers, focusing on core concepts like earth's rotation, tilt, eclipses, seasonal shifts, day-night cycles, and planetary revolution. In-depth case studies reveal significant gaps in teachers' subject matter knowledge and pedagogical content knowledge, affecting their instructional delivery. The findings show teachers struggle with abstract concepts, often using informal language and lacking scientific precision. Many demonstrated incomplete or inaccurate understanding, highlighting the need for targeted professional development to boost conceptual mastery and instructional confidence. Limited teacher knowledge can hinder learners' opportunities to develop accurate scientific understandings, emphasizing the importance of ongoing support. This research contributes to the growing evidence on the critical role of teacher knowledge in science education, particularly in complex topics like astronomy. The study's findings have implications for teacher education, professional development, and science curriculum design, underscoring the need for sustained investment in teacher support. The research highlights the importance of addressing teacher knowledge gaps to improve science education in South Africa.

Key words: Seasonal changes, Earth-sun-moon relationships, misconceptions in astronomy, astronomy education, pedagogical content knowledge.

Received 10/2025
Revised 5/2026
Accepted 6/2026

1 Introduction

The teaching and learning of astronomy have been a topic of interest in science education research for several decades. Studies have shown that students often hold misconceptions about astronomical concepts, such as the shape of the Earth, the movement of planets, and the causes of day and night (Dantic et al., 2024). Research has also highlighted the importance of teacher knowledge in astronomy education. Teachers' subject matter knowledge (SMK) and pedagogical content knowledge (PCK) are critical factors in determining the quality of instruction and student learning outcomes (Shulman, 1986). However, studies (Jansri & Ketpichainarong, 2020; Susman & Pavlin, 2020) have shown that many teachers lack the SMK and PCK needed to teach astronomy effectively, particularly in areas such as conceptual understanding and instructional design. The literature increasingly conceptualizes pedagogical content knowledge (PCK) and topic-specific PCK (TSPCK) as dynamic, learner-centered frameworks that shape student outcomes in 21st-century, digitally supported STEM education. Belayneh (2025) argues that PCK/TSPCK must integrate digital fluency and equity, a position supported by studies linking robust teacher knowledge to improved conceptual understanding, problem-solving, and cognitive engagement.

Complementary research by Dragnić-Cindrić and Anderson (2025) shows that project-based learning strengthens several dimensions of PCK—teaching orientations, curriculum knowledge, instructional strategies, and understanding student thinking—though assessment literacy remains underdeveloped. Astronomy education research provides particularly compelling evidence of the consequences of insufficient teacher knowledge: numerous studies reveal that both pre-service and in-service teachers hold persistent and widespread misconceptions about fundamental astronomy concepts (Kanli, 2014). Frede (2006) similarly found that most pre-service French elementary teachers possess non-scientific conceptions about the day/night cycle, seasonal changes, and even the spherical Earth model, despite these topics being part of the national curriculum since 1985. These misunderstandings often transfer directly to students, as demonstrated by Sule and Jawkar (2019), who reported that teacher-held misconceptions—especially regarding the Sun, Moon, and calendars—stem partly from difficulties in making logical inferences from existing knowledge. Intervention studies show partial improvement: Jansri and Ketpichainarong (2020) observed significant gains in teachers' celestial motion understanding after professional development, though misconceptions about seasons persisted, while Susman and Pavlin (2020) found that didactic astronomy games enhanced teacher knowledge and were viewed as educational and classroom-appropriate, despite teachers' continued lack of confidence with complex astronomical content.

Teachers play a critical role in shaping students' understanding of astronomy, and their own knowledge and instructional practices can have a significant impact on student learning outcomes.

The paper focused on the teacher's understanding of Earth's rotation, tilt, eclipses, seasonal shifts, day-night cycles, and planetary revolution, as well as their teaching approaches for these topics. By

examining the challenges and opportunities that teachers face in teaching astronomy, this research seeks to inform the development of targeted support and professional development programs that can help improve the quality of science education.

2 Theoretical background

Astronomy in South African schools is primarily integrated into the broader science curriculum rather than taught as a standalone subject. At the high school level, learners encounter astronomy through the Physical Sciences syllabus, where they study topics such as the solar system, stars, galaxies, and the origins of the universe. In the lower grades, within Natural Sciences, students are introduced to more foundational concepts like planetary motion, eclipses, and Earth's place in space. These lessons are often connected to mathematics, where geometry and trigonometry are applied to celestial measurements, and to history, where learners explore ancient African sky lore and cultural interpretations of the night sky. Importantly, it is the science teachers who deliver astronomy content to Physical Science students in Grades 10–12 and Natural Science students in Grades 7–9, ensuring that astronomy is taught within the framework of the existing science curriculum. This integration ensures that astronomy is not only presented as a scientific discipline but also as a subject that bridges culture, mathematics, and technology, making it both accessible and inspiring for learners across different stages of schooling.

Astronomy is a fundamental part of science education (Carli et al., 2025), providing insights into the universe and its many mysteries. However, teaching and learning astronomy can be challenging due to the abstract nature of many concepts, such as revolution, axial tilt, and eclipses. Research (Dantic et al., 2024) has shown that students often hold misconceptions about these concepts, which can be difficult to change and may persist into adulthood. Guerra-Reyes et al. (2024) have highlighted the significance of addressing misconceptions and enhancing conceptual understanding in astronomy education. For example, Şensoy and Asana (2025) highlight targeted efforts to tackle astronomy misconceptions through conceptual change texts, which have shown promise in enhancing pre-service science teachers' engagement with the subject. This suggests that tailored educational materials can be effective in promoting conceptual clarity among science teachers. In a similar vein, Nasution et al. (2025) investigated the use of Stellarium, a digital planetarium application, to enhance students' understanding of astronomy concepts in science education. Their findings indicate that technology can be a valuable tool in addressing the challenges of teaching astronomy, particularly in institutions with limited resources. Recent studies (Dantic et al., 2024; Guerra-Reyes et al., 2024; Nasution et al., 2025) underscore the need for innovative approaches to astronomy education that prioritize conceptual clarity and practical application.

Many studies (Guerra-Reyes et al., 2024; Rodrigues et al., 2025; Salimpour et al., 2024; Slater et al., 2018; Yu et al., 2010) have been done on misconceptions in astronomy, highlighting the complexities and challenges in understanding celestial concepts. Slater et al. (2018) investigated astronomy alternative conceptions in pre-adolescent students in Western Australia, revealing the prevalence of misconceptions even at a young age. Their findings underscore the need for early intervention and targeted educational strategies to address these misunderstandings. More recently, Salimpour et al. (2024) examined the mismatch between the intended astronomy curriculum content, astronomical literacy, and the astronomical universe, highlighting the disconnect between what is taught and what is relevant to understanding the universe. This disconnect can lead to a lack of depth in students' understanding of astronomy.

Further, Yu et al. (2010) explored student ideas about Kepler's laws and planetary orbital motions, shedding light on the specific areas of astronomy where students struggle. Their research demonstrates that students often hold onto misconceptions despite formal instruction, emphasizing the need for innovative teaching approaches that challenge these preconceptions. Danaia and McKinnon (2007) also identified common alternative astronomical conceptions in junior secondary science classes, suggesting that these misconceptions can be persistent and widespread. The persistence of these misconceptions highlights the need for teachers to have a deep understanding of astronomical concepts themselves, as well as effective strategies for teaching these concepts to their students.

The way astronomy is integrated into the curriculum can also impact students' understanding of astronomical concepts. This is because teachers may not always be aware of the misconceptions held by their students (Cox et al., 2016). For instance, Cox et al. (2016) found that teachers in Belgium had varying levels of awareness of students' misconceptions in astronomy, with some teachers overestimating or underestimating the prevalence of certain misconceptions. This lack of awareness can hinder teachers' ability to address these misconceptions effectively, underscoring the need for professional development opportunities that focus on both content knowledge and pedagogical content knowledge.

Rodrigues et al. (2025) analyzed the Chilean science curriculum and found that while there are opportunities to learn astronomy, these opportunities are not always fully exploited. This highlights the

importance of curriculum design and teacher support in ensuring that students receive comprehensive astronomy education.

In South Africa, primary teachers had incorrect conceptions of the Earth-Moon-Sun system and poor conceptual knowledge in basic astronomy (Govender, 2011). Similarly, K-12 STEM teachers in the US had misconceptions about the primary source of energy for stars and the eccentricity of the Earth's orbit and showed uncertainty and hesitation in their explanations (Burrows et al., 2021). Even kindergarten teachers have been found to have a lack of knowledge and alternative ideas about scientific concepts related to time, space, gravity, and light (Ampartzaki et al., 2024). A teacher's difficulty in teaching the phases of the Moon also highlights the need for improved teacher knowledge in specific astronomy concepts (Nielsen, 2014). Overall, these findings suggest that teachers often have difficulty articulating clear and accurate explanations of complex scientific concepts. The teachers' reliance on informal language and lack of scientific precision in their explanations is particularly concerning, as it can perpetuate misconceptions and hinder learners' understanding of these concepts (Ball & McDiarmid, 1990).

Literature collectively emphasizes the importance of addressing misconceptions and promoting conceptual clarity in astronomy education. By understanding the specific areas where students and teachers struggle, teachers can develop targeted interventions to improve astronomy and foster a deeper appreciation for the subject.

The impact of teacher knowledge on student learning outcomes in astronomy is significant. When teachers have a deep understanding of astronomical concepts and effective instructional strategies, they are better able to design engaging and challenging lessons that promote student understanding and motivation (Desimone, 2009). Conversely, teachers' own misconceptions and lack of knowledge can perpetuate student misconceptions and hinder learning. Therefore, it is essential to provide teachers with ongoing support and professional development opportunities (Carli et al., 2025) to enhance their knowledge and instructional practices. Desimone (2009) emphasized the need for targeted professional development programs that focus on astronomy-specific content and pedagogy. These programs should provide teachers with opportunities to develop their SMK and PCK, as well as to design and implement effective instructional strategies. By investing in teacher knowledge and instructional capacity, teachers and policymakers can work together to improve the quality of astronomy education and promote a deeper understanding of the universe among students. While the literature offers valuable insights into misconceptions research in science education, there is limited research specifically addressing astronomical misconceptions in the South African context (Guerra-Reyes et al., 2024). In this study, misconceptions are defined as scientifically inaccurate beliefs that teachers hold in place of accepted scientific explanations, while gaps refer to instances where understanding is incomplete or missing, without the presence of an alternative conception (Chi, 2013; Robinson et al., 2011; Vosniadou, 2020). Literature on astronomical misconceptions among South African science teachers builds on a robust body of evidence demonstrating persistent conceptual challenges in astronomy education. Therefore, the study explored the following research question, what misconceptions or gaps exist in teachers' conceptual understanding of Earth's rotation, tilt, eclipses, seasonal shifts, day-night cycles, and planetary revolution, and how do these affect their instructional strategies?

3 Research methodology

This study employed a qualitative research approach (Maree & Pietersen, 2016), utilizing a case study design to explore the conceptual understanding and pedagogical content knowledge (PCK) of 10 science teachers in South Africa. The ten cases had a diverse group of Grade 7 Natural Sciences teachers with a range of qualifications and experience, from mid-career teachers like Sir Lekos (25 years) and Mam Maprezz (30 years) to experienced teachers like Mam Nandi (34 years), Sir Maps (40 years), and Sir Choss (30 years). Meanwhile, early-career teachers with less experience (between 1 and 5 years) included Sir Masemos, Sir Thobs, Mam Seaps, Sir Nkomos, and Sir Bops. Each teacher has a three-year teaching credential with a focus on either physical or natural sciences. Because astronomy content is integrated into the Natural Science curriculum, teachers do not need a specialized qualification in astronomy; a three-year science education qualification provides the necessary foundation. The grade 7 learners were between 12 and 13 years old. The names used for the teachers in this study are pseudonyms and do not represent their real identities.

Purposeful sampling of ten teachers was employed to ensure the selection of information-rich participants who possess direct experience with the phenomenon under investigation, which aligns with qualitative case study traditions that prioritize depth of insight over sample size (Creswell & Poth, 2016; Patton, 2015). We used qualitative analysis because it really helped us make sense of complex, layered data, especially interviews and classroom observations. As Merriam (1988) explains, this type of anal-

ysis allows patterns, themes, and categories to naturally emerge from the data through an inductive process, which makes it especially useful for understanding how people make sense of scientific ideas in real classroom settings. In addition, case study research gains strength when it draws on multiple data sources. This process, known as triangulation, helps improve the credibility and trustworthiness of the findings, and it's something Yin (2012) strongly advocates in his methodological work. Together, the case study design and qualitative analysis give us a solid and context-rich way to explore how science teachers understand and sometimes misunderstand—astronomical concepts, as well as how those understandings influence the way they teach.

Data collection involved in-depth interviews, which were audio-recorded and transcribed verbatim. The interview protocol focused on core astronomical concepts, including revolution, axial tilt, eclipses, and seasonal changes. Thematic analysis was used to identify patterns and themes in the data, with a focus on understanding the teachers' conceptual understanding and instructional strategies. The study's qualitative design allowed for a rich exploration of the teachers' knowledge and practices, providing insights into the complexities of teaching and learning in astronomy education.

The analysis employed a qualitative case study approach, focusing on ten individual science teachers' instructional practices and conceptual understanding of astronomical topics. Data were collected through classroom observations, teacher interviews, and content-specific probes designed to elicit explanations of key concepts such as revolution, rotation, axial tilt, eclipses, and seasonal changes. Each case was analyzed thematically to identify patterns of understanding, misconceptions, pedagogical strategies, and confidence levels. Cross-case synthesis was then used to compare individual profiles, revealing recurring challenges and strengths across the cohort. This method allowed for an understanding of both content mastery and pedagogical content knowledge (PCK) in the context of teaching abstract space science concepts.

3.1 Semi-structured interviews

This study used semi-structured interviews to explore science teachers' understanding of astronomical concepts and the challenges they face in teaching the subject. The interviews allowed for in-depth probing while maintaining a consistent structure across participants (Brinkmann & Kvale, 2018). The interviews were conducted by the authors of this paper, who specialize in science education. Given that astronomy is taught as part of the broader Physical Sciences and Natural Sciences curriculum rather than as a standalone subject, the authors' expertise aligns with the curricular context in which astronomy content is delivered. Interviews lasted 45–60 minutes. They were audio-recorded with consent and provided an opportunity for teachers to reflect on their teaching methods and professional development. The interviews were essential in collecting information that influenced participants' pedagogical content knowledge (PCK) and their sense of the social world. These qualitative interviews were the most valuable source of in-depth information, focusing on the pedagogical content challenges faced by participants. The interview included the following questions:

- How do you explain the concept of *revolution* to your learners?
- How do you help learners understand the tilting of the Earth?
- What is an eclipse, and how is it related to the positions of the Sun and the Moon?
- What causes lunar and solar eclipses? Please explain how each one occurs.
- What causes day and night in relation to the Sun, the Moon, and the Earth?
- What causes seasonal changes? Please elaborate.

3.2 Classroom observations

In addition to the interviews, non-participant observations were conducted by the authors of this paper to gather further data. An observation schedule was created, focusing on PCK components, and the authors utilized both descriptive field notes and interpretive memos. Each teacher was observed in two science lessons, lasting one hour (60 minutes), with a focus on content delivery and instructional interaction. Non-participant observations reduced interference with the study, increasing its authenticity (Marshall & Rossman, 2016). The observations allowed for the collection of data through mutual responses and questions posed to participants, providing insight into their feelings, views, acceptance, and reactions (Maree & Pietersen, 2016). The observation schedule included descriptive notes, reflective notes, knowledge focus (constructivism, teacher knowledge, instructional strategies, interaction and discourse, professional development), and pedagogical content challenges.

3.3 Data analysis

Thematic analysis was used to analyze the data, following Braun and Clarke's (2021) six-phase approach (see Fig. 1).

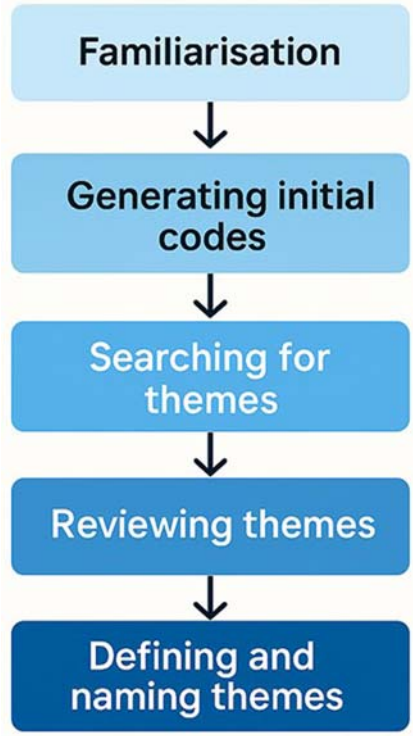

Fig. 1: Thematic analysis (Braun & Clarke, 2021)

This method was chosen for its flexibility and its suitability for exploring teachers' experiences and conceptual understandings in depth. The analysis began with a period of intensive familiarization, during which the researchers read and re-read each transcript to gain a sense of the overall patterns emerging from the teachers' responses. At this stage, early observations began to surface—particularly regarding the level of confidence and clarity with which teachers described astronomy concepts.

During the coding phase, the researchers worked systematically through the transcripts, generating initial codes both inductively and with reference to the research questions. For instance, when one teacher remarked, "To be honest, even we as teachers sometimes need clearer explanations ourselves. It's hard... to confidently teach these concepts when our own understanding isn't deep enough," this statement was coded as an example of limited content knowledge and reduced confidence in teaching astronomy. Similarly, when another teacher explained an eclipse as "the falling of the shadow of the Earth either on the sun or the moon," the coding captured the presence of misconceptions about celestial mechanics. These early codes highlighted recurring uncertainties and conceptual gaps in teachers' explanations.

As coding continued, patterns began to emerge. Teachers frequently conflated basic astronomical concepts such as rotation and revolution. One teacher, attempting to explain axial tilt, described, "When you tilt neh!... there is an imaginary line... So when we tilt, uummhh when you spin neh! It is no longer movement we call revolution." Statements like this were coded as instances of conceptual confusion and inconsistent use of scientific terminology. Other teachers expressed uncertainty about eclipses, with comments such as, "I have not revised that, but I think it is when the moon shows the dark side..." or "...the eclipse I do not remember very well... I guess those notions I forgot." These responses were coded as indicators of limited recall and the need for revision of fundamental astronomy concepts.

The third phase of the analysis involved grouping related codes into broader, preliminary themes. For example, codes relating to conceptual confusion, incorrect explanations of astronomical events, and the misuse of terminology began to cluster under the emerging theme of "persistent misconceptions". Likewise, statements expressing uncertainty, lack of revision, or insufficient confidence were grouped into a developing theme concerning teachers' limited conceptual mastery and confidence.

These themes were refined through an iterative review process, where the researchers revisited both the coded data and the full dataset to ensure coherence and internal consistency. During this stage, themes were consolidated, clarified, and strengthened. For example, individual categories relating to teachers' uncertainty, forgotten content, and low confidence were merged into a single, more comprehensive

theme reflecting broader content-knowledge limitations. Similarly, misconceptions related to eclipses, seasons, and planetary motion were unified under a theme addressing conceptual inaccuracies in teachers' understanding of astronomy.

The fifth phase of analysis involved defining and naming the final themes, ensuring that each theme captured the essence of the coded data while remaining analytically distinct. Key themes included teachers' persistent misconceptions of core astronomy concepts, limited confidence and content mastery, and inconsistent use of scientific terminology. These themes reflected the most prominent challenges teachers experienced when attempting to teach astronomy.

Finally, the themes were woven into a coherent analytic narrative during the reporting phase. The teachers' own words provided rich, illustrative examples of the conceptual difficulties they faced, from describing eclipses inaccurately to confusing fundamental celestial motions. Through this process, the analysis revealed not only the gaps in teachers' astronomy knowledge but also the pressures and uncertainties that shaped their instructional practices. The findings ultimately informed recommendations for targeted professional development aimed at strengthening teachers' conceptual understanding and improving their confidence in delivering astronomy content.

4 Findings

This study presents a case analysis of ten science teachers—Mam Nandi, Sir Lekos, Mam Maprezz, Sir Masemos, Sir Thobs, Sir Maps, Sir Choss, Mam Seaps, Sir Nkomos, and Sir Bops—focused on their instructional approaches and conceptual understanding of astronomical topics. The analysis of the cases studies identified common strengths, challenges, and areas for professional development in teaching abstract astronomy concepts. The findings are presented as a thematic analysis organized into three principal areas: teachers' conceptual understanding of astronomical concepts, their pedagogical approaches, and individual teacher profiles.

4.1 Teacher's conceptual understanding of astronomy concepts

Table 1 presents a thematic summary of teachers' conceptual understanding across five core astronomical topics: revolution, rotation, axial tilt, eclipses, and seasonal changes. Table 1 highlights common strengths observed among the ten participating teachers, as well as recurring conceptual challenges that emerged during classroom observations and interviews. While many teachers demonstrated a foundational grasp of certain concepts—such as linking rotation to day and night or recognizing the role of revolution in seasonal changes, significant gaps in depth, clarity, and scientific accuracy were evident. The findings highlight that teaching about the relationship between the Sun, Moon, and Earth presents difficulties for teachers. As one of the teachers (Mam Seaps) reflected: "To be honest, even we as teachers sometimes need clearer explanations ourselves. It's hard or difficult to confidently teach these concepts when our own understanding isn't deep enough." This admission underscores the central issue identified in the study—teachers' own conceptual uncertainties can undermine their confidence and clarity in the classroom.

Table 1: Conceptual understanding of astronomy

Concept Area	Common Strengths	Common Challenges
Revolution	Most teachers could define revolution in basic terms.	Explanations often lacked clarity, depth, or were confused with rotation.
Rotation	Several teachers correctly linked rotation to day/night.	Many confused rotation with tilt or gave hesitant, vague explanations.
Axial Tilt	A few teachers attempted to use models or gestures to explain tilt.	Frequently misunderstood, omitted, or conflated with other concepts.
Eclipses	Some awareness of solar/lunar roles.	Widespread misconceptions, vague or incorrect explanations, and lack of revision.
Seasonal Changes	General link to revolution was common.	Axial tilt often omitted; explanations were oversimplified or hesitant.

Teachers' explanations of eclipses often revealed significant conceptual gaps and imprecise use of scientific language. For example, when Sir Lekos was asked what causes lunar and solar eclipses, the teacher responded: "It is caused by the falling of the shadow of the earth either on the sun or the moon." This explanation reflects a partial understanding of shadow phenomena but conflates the distinct mechanisms of solar and lunar eclipses. Scientifically, a solar eclipse occurs when the Moon passes between the Earth and the Sun, casting a shadow on Earth, while a lunar eclipse occurs when the Earth passes between the Sun and the Moon, casting its shadow on the Moon.

The teacher's phrasing, "falling of the shadow of the earth either on the sun or the moon", not only misrepresents the geometry of these events but also risks reinforcing misconceptions among learners. Such oversimplification obscures the critical role of alignment between celestial bodies and fails to distinguish between the two types of eclipses.

This overview provides a snapshot of the collective conceptual landscape, serving as a basis for identifying targeted areas for professional development.

Table 1 summarizes the common strengths and challenges observed across the ten teachers in their understanding and instruction of key astronomical concepts. Most teachers were able to provide basic definitions of revolution and correctly associated rotation with the occurrence of day and night. However, their explanations often lacked depth and precision, with frequent confusion between rotation and revolution. Teachers' explanations of astronomical concepts often revealed gaps in depth and scientific precision. For example, when Mam Nandi was asked how they explained the concept of *revolution* to learners, Mam Nandi responded: "Eeeeeeh yah, revolution is the travelling of the planet or the bodies because there are also the asteroids of bodies around the sun on their orbit which is as simple as a path." While this response captures the idea of orbital motion, the explanation is vague, conflates different celestial bodies, and lacks the clarity needed to distinguish Earth's revolution around the Sun from other orbital phenomena. Axial tilt was occasionally demonstrated using models or gestures, yet its scientific significance—particularly in explaining seasonal changes—was frequently misunderstood or omitted. While there was partial awareness of the roles of the sun and moon in eclipses, widespread misconceptions and vague explanations were evident. Seasonal changes were generally linked to the earth's revolution, but the critical role of axial tilt was often neglected, resulting in oversimplified accounts. These patterns highlight the need for strengthened content knowledge and pedagogical clarity in teaching abstract astronomical phenomena.

4.2 Pedagogical approaches of teachers

The data revealed that several teachers incorporated physical models and gestures when teaching astronomical concepts, which may form part of a broader constructivist strategy if combined with learner-centered activities. Teachers like Sir Thobs, Sir Masemos, and Sir Choss demonstrated efforts to make abstract ideas more tangible through visual aids; however, these attempts were frequently undermined by conceptual confusion and vague terminology, which limited their instructional effectiveness. Additionally, the use of informal or imprecise language was a recurring issue across multiple cases, often obscuring scientific meaning and increasing the risk of learner misconceptions. Compounding these challenges, teachers such as Mam Seaps, Sir Maps, and Sir Bops openly acknowledged not revising key content areas, which contributed to hesitant delivery and incomplete explanations. These findings highlight the interplay between pedagogical strategies, language clarity, and content preparedness in shaping the quality of science instruction.

4.3 Individual teacher profiles

Table 2 presents a summary of individual teacher profiles, highlighting both their instructional strengths and areas requiring development in the context of teaching astronomical concepts. Each teacher demonstrated unique approaches and varying levels of conceptual understanding, with some showing promise in constructivist strategies and contextual reasoning, while others struggled with scientific clarity and pedagogical confidence. This comparative overview provides a concise snapshot of how each teacher engaged with core astronomy topics, offering insight into the diversity of teaching practices and the specific support needed to enhance their pedagogical content knowledge (PCK). To ensure transparency and consistency in how the "areas for development" were identified, the teachers' explanations were analyzed using a set of clearly defined criteria.

These criteria focused on four key dimensions of conceptual understanding: scientific accuracy, referring to whether the explanation aligned with accepted astronomical principles; conceptual coherence, which considered the teacher's ability to correctly relate concepts such as rotation, revolution, axial tilt, seasons, and eclipses; scientific language and precision, assessing the degree of clarity and correctness in terminology used; and depth of reasoning, which evaluated whether the explanation moved beyond surface-level descriptions to demonstrate a more developed conceptual framework. These indicators guided the coding process and allowed for systematic differentiation between minor lapses in clarity and more substantive conceptual misunderstandings. The resulting "areas for development" summarized in Table 2 therefore reflect consistent patterns in the teachers' responses and highlight specific aspects of astronomical content knowledge that require further strengthening.

Table 2 indicates a comparative overview of individual teacher profiles, highlighting both instructional strengths and areas for development in teaching astronomical concepts. Several teachers, such as Mam

Table 2: Individual teacher profiles

Teacher	Strengths	Areas for development	Interview except examples
Mam Nandi	Creative teaching approach; linked seasons to revolution.	Misconceptions about eclipses; confused rotation with tilt; lacked scientific precision.	*Revolution is the travelling of the planet or the bodies because there are also the asteroids of bodies around the sun on their orbit, which is as simple as a path.*
Sir Lekos	Used gestures and visual aids; partial understanding of eclipses.	Simplistic explanations; confused rotation and revolution; omitted axial tilt.	*Revolution is the movement of Earth around the sun. When I was describing rotation I used the word "spinning," and then revolution is moving around something. Yes, the learners were able to understand the two concepts of rotation and revolution.*
Mam Maprezz	Strong conceptual grasp; contextualized explanations; integrated leap years.	Needs refinement in scientific language and clarity.	*Revolution occurred when the Earth moves around the sun for three hundred and sixty-five days (365) and six hours or a quarter day (1/4) in one year, which is 365 1/4 days. The moon revolves around the Earth for twenty-eight days (28) in one month, and the Earth rotates around its own axis for twenty-four hours (24 hrs) in one day. This rotation happens at an angle of 23.5° through the process called tilting.*
Sir Mase-mos	Proposed use of models; constructivist intent.	Tentative explanations; vague terminology; admitted lack of revision.	*Revolution eehh... eehh... I understand it as a concept that deals with circulation of things, things that come time and again and keep on coming back again. That is my basic understanding of the word revolution. And I understand that the Earth revolves around the sun, the sun is stable there, and other objects also revolve around the sun, which is why we say the sun is at the center of the solar system.*
Sir Thobs	Used improvised models; constructivist approach.	Lacked depth in explanations; misconceptions about eclipses.	*eish is it not related to the moon? Uummhh but I guess the one for the sun is by the time when the moon and the Earth are direct and the other one, I guess it happened once in my lifetime and it just closed the sun and then the sun became big dark, and the moon would have a shadow and we get an eclipse.*
Sir Maps	Attempted learner-centered strategies.	Confused axial tilt; vague eclipse explanations; misunderstood day/night cycle.	*The eclipse we talk of, where we see half of the moon being dark or the part of the sun being dark. Eish, I have not done revision of these eclipses, I forgot.*
Sir Choss	Enthusiastic; proposed hands-on activities.	Confused tilt with rotation; vague eclipse explanations.	*When you tilt neh! There is a pattern because you cannot divorce one from the other. There is an imaginary line i.e., axis. Let me come first to the sun: when we tilt it is like you spin, and there is this called imaginary line that cuts the Earth in between from the northern hemisphere and southern hemisphere. So when we tilt, uummhh when you spin neh! It is no longer movement we call revolution.*
Mam Seaps	Correctly linked rotation to day/night.	Hesitant delivery; conflated tilt and rotation; vague on eclipses and seasons.	*In terms of eclipse neh! Eish eehh... I have not revised that, but I think it is when the moon shows the dark side and when the sun is dim or dark because of the movement of the Earth in between.*
Sir Nkomos	Partial understanding of day/night.	Fragmented knowledge; lacked scientific terminology; confused eclipse types.	*Eclipse I am not sure that I am gonna say it correctly but, if the moon is on the other side and the sun is on the other side then the Earth is between them. Yah, let me start with the one when the Earth is between the moon and the sun. Ok, the lunar eclipse is when the Earth obstructs the moon, and when the Earth is also obstructing the sun, and both appear to be black or reddish in color.*
Sir Bops	Willingness to engage despite uncertainty.	Severe content gaps; lacked clarity and confidence.	*... even the eclipse I do not remember it very well. I guess lunar has to do with the moon, where the solar is relatively to the sun. Eish, those notions I forgot, and I guess they needed a thorough revision.*

Maprezz and Mam Nandi, demonstrated promising strengths in conceptual understanding and contextual reasoning, with Mam Maprezz notably integrating leap years into her explanations. Others, like Sir Thobs, Sir Masemos, and Sir Choss, showed constructivist intent through the use of models and hands-on activities, reflecting efforts to make abstract concepts more tangible. However, across the cohort, significant challenges were evident. Misconceptions about eclipses, confusion between rotation and axial tilt, and vague or hesitant explanations were recurring issues. Teachers such as Sir Maps, Sir Nkomos, and Sir Bops exhibited critical gaps in scientific clarity and pedagogical confidence, often compounded by limited content revision. The table underscores the need for targeted professional development to address both conceptual misunderstandings and instructional delivery, ensuring that teachers are equipped to support accurate and engaging astronomy education.

The analysis revealed three prominent challenges that hindered effective astronomy instruction across the teacher cohort (Table 2). First, conceptual confusion was widespread, with many teachers struggling to distinguish between rotation, revolution, and axial tilt—leading to inaccurate explanations of phenomena such as day and night or seasonal changes. Second, a lack of scientific precision was evident in the delivery of content; explanations were frequently vague, hesitant, or scientifically incorrect, which risks reinforcing misconceptions among learners. Third, gaps in pedagogical content knowledge (PCK) were apparent, as teachers often found it difficult to translate their understanding into clear, structured, and learner-friendly teaching strategies. These challenges collectively highlight the need for targeted professional development that strengthens both conceptual mastery and instructional competence in space science education.

5 Discussion

The cases of Mam Nandi, Sir Lekos, Mam Maprezz, Sir Masemos, Sir Thobs, Sir Maps, Sir Choss, Mam Seaps, Sir Nkomos, and Sir Bops collectively highlight the complexities and challenges science teachers face in teaching astronomical concepts. A common thread among these cases is the struggle with precision, clarity, and depth in explaining abstract concepts such as eclipses, axial tilt, rotation, and revolution. For instance, Mam Nandi and Sir Lekos demonstrated fundamental misunderstandings about eclipses, while Sir Maps and Sir Choss confused the causes of day and night, attributing them to Earth's tilting rather than rotation.

These cases reflect broader trends identified in recent literature. Dantic et al. (2024) found that pre-service science teachers hold persistent misconceptions about celestial phenomena, particularly eclipses and planetary motion. Yu et al. (2010) emphasized that such misconceptions are deeply ingrained and require deliberate instructional strategies to correct. Cox et al. (2016) further noted that teachers often lack awareness of student misconceptions and struggle to adapt instruction accordingly, a challenge mirrored in the fragmented and hesitant explanations given by Sir Nkomos and Mam Seaps.

The cases also reveal significant gaps in teachers' conceptual understanding, with many displaying uncertainty. Mam Seaps, for example, showed partial grasp of revolution but lacked confidence, using tentative language. Sir Bops's struggles were even more pronounced, admitting to not revising key concepts and guessing about the causes of lunar and solar eclipses. Sir Nkomo's responses reflected incomplete and loosely framed conceptual knowledge, with fragmented, hesitant, and imprecise explanations. These gaps in teacher knowledge can directly hinder learner's opportunities to develop accurate scientific understandings of complex natural phenomena.

Recent studies underscore the importance of pedagogical content knowledge (PCK) in addressing these challenges. Belayneh (2025) argues for a dual framework of general and topic-specific PCK (TSPCK), which is crucial for aligning instruction with student needs and digital fluency. Dragnić-Cindrić and Anderson (2025) found that pre-service teachers often develop uneven PCK, with strong teaching orientations but weak understanding of student cognition and assessment issues evident in the cases of Mam Seaps and Sir Nkomos. Nielsen (2014) suggests that integrating student feedback into lesson planning can enhance enacted PCK, helping teachers refine their instructional approaches.

The cases strongly support the necessity of targeted, topic-specific training programs to strengthen teachers' basic knowledge and instructional competence in space science topics. With continuous professional development, teachers like Mam Seaps, Sir Nkomos, and Sir Bops could improve their content mastery and confidence, enhancing their ability to engage learners and correct misconceptions. Rehman et al. (2025) advocate for such targeted professional development, emphasizing its role in addressing conceptual gaps and pedagogical inconsistencies.

By providing teachers with the necessary support and resources, they can develop confidence and expertise to provide accurate and engaging instruction, ultimately enhancing learner understanding and achievement in astronomy. Moreover, such support can help teachers move beyond mere transmission of facts and instead facilitate meaningful learning experiences that promote scientific literacy and critical

thinking. The cases also highlight the importance of pedagogical content knowledge in teaching astronomy. Teachers need not only to possess a deep understanding of the subject matter but also to be able to convey complex concepts in a clear and concise manner. By focusing on both content knowledge and pedagogical skills, professional development programs can help teachers become more effective instructors and foster a deeper appreciation of astronomy among their learners (Belayneh, 2025).

The results of this study are consistent with existing research on teachers' understanding of astronomy concepts. The study's findings align with Frede's (2006) research, which suggests that teachers often hold non-scientific understandings of the astronomical concepts they teach. This is further supported by studies such as Kanli (2014), which found extensive misconceptions among pre-service and in-service science teachers about basic astronomy concepts like seasons, moon phases, and the sun's position. Similarly, Stears et al. (2011) reported limited conceptual change in teachers' understanding after a module, attributing this to ineffective instructional strategies. Sule and Jawkar (2019) also highlighted significant skill deficits in imparting correct astronomical knowledge due to information gaps and poor logical reasoning. These findings collectively suggest that teachers' misconceptions and skill deficits in astronomy are persistent issues that require targeted professional development and support. The current study's results reinforce the need for teachers to address these gaps in knowledge and understanding to improve the teaching and learning of astronomy.

The findings of this study highlight the significant challenges that science teachers face in developing a deep understanding of astronomical concepts, such as revolution, axial tilt, eclipses, and seasonal changes. These results are consistent with previous research that has shown that many teachers struggle to articulate clear and accurate explanations of complex scientific concepts (Kanli, 2014; Stears et al., 2011; Sule & Jawkar, 2019). Research consistently shows that many teachers struggle to clearly explain complex scientific concepts, particularly in astronomy. Studies have documented specific misconceptions among teachers, such as believing that seasons occur due to the Earth's distance from the Sun rather than its axial tilt, as found in a study of 45 secondary science teachers in Thailand where 42% held this misconception (Jansri & Ketpichainarong, 2020). This lack of understanding is not limited to secondary teachers, as primary teachers in Slovenia struggled with switching between heliocentric and geocentric systems and understanding frames of reference (Susman & Pavlin, 2020).

The findings of this study also underscore the importance of pedagogical content knowledge (PCK) in teaching astronomy. The teachers' ability to design effective instructional strategies and assessments depends on their understanding of the subject matter and their ability to communicate complex concepts in a clear and concise manner (Shulman, 1986). However, the results of this study suggest that many teachers lack the PCK needed to teach astronomy effectively, which can have a negative impact on learners' understanding and engagement with the subject.

The need for targeted professional development programs that focus on strengthening teachers' conceptual mastery and instructional confidence is clear. Research has shown that effective professional development programs can have a positive impact on teachers' knowledge and practices, leading to improved student outcomes (Desimone, 2009). However, to design such programs effectively, it is essential to first identify the specific misconceptions and gaps in teachers' understanding of key astronomical concepts—such as Earth's rotation, axial tilt, eclipses, seasonal shifts, day-night cycles, and planetary revolution—and examine how these shortcomings shape their instructional strategies. By investing in teacher support and development, teachers and policymakers can work together to improve science education and promote a deeper understanding of the natural world among learners.

6 Limitations of the study

While this case analysis provides valuable insights into science teachers' understanding and instructional approaches to astronomical concepts, several limitations must be acknowledged. First, the study draws on in-depth data from ten teachers; therefore, the findings are intended to provide contextualized insights rather than broad generalizations. Second, data were primarily derived from classroom observations and semi-structured interviews, which may be influenced by participant bias and performance anxiety. Third, the study did not include learner perspectives or achievement data, which could have provided a more comprehensive view of instructional impact. Additionally, variations in teaching experience, school resources, and curriculum exposure were not fully controlled, potentially affecting the consistency of findings. Future research could address these limitations by incorporating larger, more diverse samples, triangulating data sources, and examining learner outcomes alongside teacher practices.

7 Conclusion

Case analysis has illuminated the multifaceted challenges that Grade 7 Natural Sciences teachers encounter when teaching astronomical concepts. The study's central research question—examining misconceptions or gaps in teachers' conceptual understanding of Earth's rotation, axial tilt, eclipses, seasonal shifts, day–night cycles, and planetary revolution, and how these affect instructional strategies—was reflected in the findings. Based on the analysis of ten Grade 7 Natural Sciences teachers, some teachers across varying levels of experience demonstrated strengths in learner engagement and partial conceptual clarity; however, the majority faced challenges with scientific precision, pedagogical confidence, and effectively communicating abstract astronomical ideas. Recurring misconceptions, particularly around rotation, revolution, tilt, and eclipses, were evident among both early-career teachers and seasoned teachers with decades of practice. These findings underscore that teaching experience alone does not guarantee conceptual mastery, highlighting the urgent need for targeted, topic-specific professional development.

Importantly, the study also points to the significance of the range of subjects studied during the teacher education program. Because Natural Sciences is a broad learning area—encompassing life sciences, physical sciences, earth sciences, and technological content, the depth and balance of exposure teachers receive during training becomes critical. Many pre-service programs offer limited coursework in astronomy compared to other science domains, which may contribute to the persistent misconceptions identified in this study. Strengthening teacher preparation by ensuring that all major science strands, including astronomy, receive adequate conceptual and pedagogical attention would better equip teachers to navigate the interdisciplinary demands of the curriculum.

Strengthening teachers' pedagogical content knowledge (PCK), refining their use of scientific language, and equipping them with strategies to address learner misconceptions are essential steps toward improving astronomy education. By investing in continuous support and structured training, teachers and policymakers can ensure that teachers—regardless of career stage—are better prepared to foster accurate understanding, correct misconceptions, and inspire curiosity about space science among learners aged 12 to 13.

Ultimately, the study's findings carry significant implications for teacher education, professional development, and curriculum design. Prioritizing teacher knowledge, the breadth of subject exposure in pre-service programs, and instructional capacity will be vital in advancing science education and promoting a deeper, more accurate understanding of the natural world among learners.

References

Ampartzaki, M., Tassis, K., Kalogiannakis, M., Pavlidou, V., Christidis, K., Chatzoglidou, S., & Eleftherakis, G. (2024). Assessing the initial outcomes of a blended learning course for teachers facilitating astronomy activities for young children. *Education Sciences*, *14*(6), 606. https://doi.org/10.3390/educsci14060606

Ball, D. L., & McDiarmid, G. W. (1990). The subject-matter preparation of teachers. In R. R. Hawley & J. D. Hawley (Eds.), *The handbook of research on teacher education* (pp. 437–449). Macmillan. https://files.eric.ed.gov/fulltext/ED310084.pdf

Belayneh, K. D. (2025). Reframing pedagogical content knowledge and topic-specific pedagogical content knowledge as dual frameworks for teaching and learning. *Pedagogical Research*, *10*(3), Article em0247. https://doi.org/10.29333/pr/16854

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. Sage Publications.

Brinkmann, S., & Kvale, S. (2018). *Doing interviews*. Sage Publications. https://doi.org/10.4135/9781529716665

Burrows, A. C., Borowczak, M., Myers, A., Schwortz, A. C., & McKim, C. (2021). Integrated STEM for teacher professional learning and development: "I need time for practice". *Education Sciences*, *11*(1), 21. https://doi.org/10.3390/educsci11010021

Carli, M., Leonardi, A. M., Ciroi, S., & Pantano, O. (2025). Teaching physics through astronomy in secondary school: The Asiago Teachers' Network on Astrophysics. *Journal of Physics: Conference Series*, *2950*(1), 012034. https://doi.org/10.1088/1742-6596/2950/1/012034

Chi, M. T. H. (2013). Two kinds and four sub-types of misconceived knowledge, ways to change it, and the learning outcomes. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (2nd ed.) (pp. 49–70). Routledge. https://doi.org/10.4324/9780203154472.CH3

Cox, M., Steegen, A., & De Cock, M. (2016). How aware are teachers of students' misconceptions in astronomy? A qualitative analysis in Belgium. *Science Education International*, *27*(2), 277–300. https://files.eric.ed.gov/fulltext/EJ1104665.pdf

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches.* SAGE publications.

Danaia, L., & McKinnon, D. (2007). Common alternative astronomical conceptions encountered in junior secondary science classes: Why is this so? *Astronomy Education Review*, *6*, 32–53. https://doi.org/10.3847/AER2007017

Dantic, M. J., Molnar, J. M. T., Alves, M. C., Calma, M., & Pascual, R. A. (2024). Misconceptions of the science education freshmen students towards astronomy. *Galaxy International Interdisciplinary Research Journal*, *12*(4). https://internationaljournals.co.in/index.php/giirj/article/view/5508

Dragnić-Cindrić, D., & Anderson, J. L. (2025). Developing pre-service teachers' pedagogical content knowledge: Lessons from a science methods class. *Education Sciences*, *15*(7), Article 860. https://doi.org/10.3390/educsci15070860

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. https://doi.org/10.3102/0013189X0833114

Frede, V. (2006). Pre-service elementary teacher's conceptions about astronomy. *Advances in Space Research*, *38*(10), 2237–2246. https://doi.org/10.1016/j.asr.2006.02.017

Govender, N. (2011). South African primary school teachers' scientific and indigenous conceptions of the Earth-Moon-Sun system. *African Journal of Research in Mathematics, Science and Technology Education*, *15*(2), 154–167. https://doi.org/10.1080/10288457.2011.10740709

Guerra-Reyes, F., Guerra-Dávila, E., Naranjo-Toro, M., Basantes-Andrade, A., & Guevara-Betancourt, S. (2024). Misconceptions in the learning of natural sciences: A systematic review. *Education Sciences*, *14*(5), 497. https://doi.org/10.3390/educsci14050497

Jansri, S., & Ketpichainarong, W. (2020). Investigating in-service science teachers' conceptions of astronomy, and determine the obstacles in teaching astronomy in Thailand. *International Journal of Educational Methodology*, *6*(4), 745–758. https://doi.org/10.12973/ijem.6.4.745

Kanli, U. (2014). A study on identifying the misconceptions of pre-service and in-service teachers about basic astronomy concepts. *Eurasia Journal of Mathematics, Science and Technology Education*, *10*(5), 471–479. https://doi.org/10.12973/eurasia.2014.1120a

Maree, K., & Pietersen, J. (2016). *First steps in research.* Van Schaik Publishers.

Marshall, C., & Rossman, G. B. (2016). *Designing qualitative research* (7th ed.). SAGE Publications.

Merriam, S. B. (1988). *Case study research in education: A qualitative approach.* Jossey-Bass.

Nasution, L. A., Khairiah, K., Siregar, J., Destini, R., Kurniawan, C., & Binti Rusli, R. (2025). Effectiveness of Stellarium application in enhancing student's understanding: Astronomy concepts in physics education. *Jurnal Eduscience*, *12*(4), 957–966. https://jurnal.ulb.ac.id/index.php/eduscience/article/view/7274

Nielsen, B. L. (2014). Students' annotated drawings as a mediating artefact in science teachers' professional development. *NorDiNa: Nordic Studies in Science Education*, *10*(2), 162–175. https://www.ucviden.dk/ws/portalfiles/portal/140188127/NorDiNa_annotated_drawings.pdf

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). SAGE Publications. https://lccn.loc.gov/2014029195

Rehman, N., Huang, X., Mahmood, A., Zafeer, H. M. I., & Mohammad, N. K. (2025). Emerging trends and effective strategies in STEM teacher professional development: A systematic review. *Humanities and Social Sciences Communications*, *12*, 32. https://doi.org/10.1057/s41599-024-04272-y

Robinson, K. A., Saldanha, I. J., & McKoy, N. A. (2011). Development of a framework to identify research gaps from systematic reviews. *Journal of Clinical Epidemiology*, *64*(12), 1325–1330. https://doi.org/10.1016/j.jclinepi.2011.06.009

Rodrigues, L., Meneses, A., Montenegro, M., & Cortès, C. (2025). Direct and indirect opportunities to learn astronomy within the Chilean science curriculum. *International Journal of Science and Mathematics Education*, *23*(1), 161–191. https://doi.org/10.1007/s10763-024-10459-1

Salimpour, S., Fitzgerald, M., & Hollow, R. (2024). Examining the mismatch between the intended astronomy curriculum content, astronomical literacy, and the astronomical universe. *Physical Review Physics Education Research*, *20*. https://doi.org/10.1103/PhysRevPhysEducRes.20.010135

Şensoy, Ö., & Asana, Y. Y. (2025). Prospective teachers' interest in astronomy: The effect of conceptual change texts in astronomy. In Cheng-Yao Lin & Li Sun (Eds.), *Diversity, Equity, and Inclusion for Mathematics and Science Education: Cases and Perspectives* (pp. 97–144). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-0345-1.ch004

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X01500200

Slater, E. V., Morris, J. E., & McKinnon, D. (2018). Astronomy alternative conceptions in pre-adolescent students in Western Australia. *International Journal of Science Education*, *40*(17), 2158–2180. https://doi.org/10.1080/09500693.2018.1522014

Stears, M., James, A., & Good, M. A. (2011). Teachers as learners: A case study of teachers' understanding of astronomy concepts and processes in an ACE course. *South African Journal of Higher Education*, *25*(3), 568–582. https://hdl.handle.net/10520/EJC37690

Sule, A., & Jawkar, S. (2019). Teacher's misconception in curricular astronomy. In *EPJ Web of Conferences* (Vol. 200, p. 01012). EDP Sciences. https://doi.org/10.1051/epjconf/201920001012

Susman, K., & Pavlin, J. (2020). Improvements in teachers' knowledge and understanding of basic astronomy concepts through didactic games. *Journal of Baltic Science Education*, *19*(6), 1020–1033. https://www.ceeol.com/search/article-detail?id=944960

Vosniadou, S. (2020). Students' misconceptions and science education. In Li-fang Zhang (Ed.), *Oxford Research Encyclopedia of Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.965

Yin, R. K. (2012). A (very) brief refresher on the case study method. *Applications of case study research*, *3*, 3–20.

Yu, K. C., Sahami, K., & Denn, G. (2010). Student ideas about Kepler's laws and planetary orbital motions. *Astronomy Education Review*, *9*(1), 010108–010117. https://doi.org/10.3847/AER2009069

How to conduct inquiry: Planning skills of lower secondary school students

Adam Nejedlý[1,*], Karel Vojíř[1]

[1] Faculty of Education, Charles University, Magdalény Rettigové 4, 116 39 Praha 1, Czech Republic; adam.nejedly@pedf.cuni.cz

The inquiry-based approach is the fundamental way of acquiring new knowledge in science. It is therefore essential to incorporate it into science education. A key principle is bridging the gap between asking questions and final answers, which requires an appropriate inquiry procedure. This study examined the planning skills of lower secondary school students when solving biology tasks based on an inquiry-based approach. Sets of solutions to six tasks from 51 Czech students were analysed. Closed and open coding based on research design aspects derived from the EDAT framework was used. The results show that most students recognized the need for an empirical approach and frequently identified observed variables and instruments. However, aspects of the inquiry process, such as controlling variables, replication, data analysis, and reducing uncertainty, were rarely included or were omitted. Differences were also found between tasks with descriptive and causal problems and across task contexts. The results suggest that inquiry-based learning needs more explicit support for the development of planning skills, with a focus on the complexity of research procedures. It also appears necessary to address these skills across various thematic units.

Key words: inquiry-based learning, planning skills, biology education.

Received 3/2026
Revised 5/2026
Accepted 6/2026

1 Introduction

Climate change, biodiversity loss, the spread of misinformation, and rapid technological advancement are just some of the complex global challenges facing society today (OECD, 2025; World Economic Forum, 2026). Solving these problems requires the active participation of citizens who are able to understand science, technology, and mathematics, evaluate them critically, and apply them in both individual and collective decision-making (Mocanu et al., 2025). In this context, science literacy is considered a key outcome of science education (Elhai, 2023; OECD, 2023b).

Empirical data suggest that achieving these goals remains challenging (Kranz et al., 2023). The PISA 2022 results show that a significant proportion of students in European Union countries do not meet the basic level of science literacy; in the Czech Republic, more than one-fifth of 15-year-old students fall below this level (OECD, 2023a). These findings indicate that some students have limited ability to apply science knowledge to solve real-world problems and to critically evaluate scientific information and distinguish warranted trust in science from either uncritical acceptance or sceptical rejection of expert knowledge (Baltikian et al., 2024; Osborne & Allchin, 2024; Teig, 2024).

Research has repeatedly highlighted the difficulties students face when solving tasks that require a higher level of inquiry-related skills, particularly in identifying and operationalizing variables (Čipková et al., 2025), formulating research questions (Nejedlý & Vojíř, 2022) and hypotheses (Ješková et al., 2022) or supporting conclusions with relevant evidence (Nejedlý & Vojíř, 2025).

Planning an investigation is commonly regarded as one of the core components of scientific inquiry and scientific literacy. In assessment frameworks such as PISA, planning, interpreting evidence, and evaluating scientific procedures are conceptualised as related dimensions of students' scientific competence. Nevertheless, the relationship between these dimensions should not be treated as self-evident. It requires empirical examination, particularly in the context of inquiry-based tasks implemented in real educational settings. For this reason, a detailed analysis of students' planning skills can provide important evidence about how lower secondary school students approach the methodological process of answering a research question. Such analysis may also help clarify the extent to which students' planning of an investigation supports, or is associated with, their later interpretation and evaluation of evidence. The aim of the research was to determine the skills of lower secondary school students in planning the process of inquiry for solving biology tasks. This aim was specified using the following research questions:

1. What aspects do students include when planning a methodological approach to answering a research question while working on inquiry-based tasks in biology?
2. Do students include different aspects when planning a methodological approach to answering descriptive and causal research questions while working on inquiry-based tasks in biology?
3. What are the typical student solutions when planning a methodological procedure for answering research questions in model inquiry-based tasks in biology?

2 Theoretical background

2.1 Science literacy as a key outcome of science education

According to the OECD (2023b) conceptual framework, science literacy encompasses the ability to (a) explain natural phenomena scientifically, (b) design and evaluate research procedures and critically interpret data and evidence, and (c) locate, assess, and use scientific information in decision-making and action (OECD, 2023b). Science literacy thus goes beyond the acquisition of factual knowledge and includes an understanding of the scientific process, the nature of science, the role of evidence and uncertainty, distinguishing between correlation and causation, and applying knowledge in new contexts (NASEM, 2016; Osborne & Allchin, 2024). Recent work further emphasises that science education should prepare students not only to evaluate evidence directly, but also to recognise the limits of their own expertise and to make informed judgements about the credibility of scientific sources (Osborne & Allchin, 2024). Osborne and Allchin (2024) argue that citizens are inevitably epistemically dependent on scientific experts and therefore need to develop informed epistemic trust, epistemic humility, and the competencies of a "competent outsider": the ability to assess who is legitimately speaking for science, whether a relevant expert consensus exists, and how scientific claims have been produced, reviewed, communicated, and potentially distorted in public contexts.

2.2 Skills in the inquiry process

The emphasis is on the limitations in students' skills connected with inquiry, which constitute a key component of science literacy and encompass a set of cognitive and metacognitive processes associated with conducting and interpreting scientific investigations (National Research Council, 2012; Osborne & Dillon, 2008). In the literature, skills connected with inquiry are often conceptualised as a set of interrelated processes that encompass the various stages of the inquiry cycle (Pedaste et al., 2015). These skills can be further divided into planning skills, analytical skills, and interpretive skills (Fugarasti et al., 2019; Pedaste et al., 2021). These components are interrelated, and their development depends on both a sufficient knowledge base and opportunities for active participation in research activities (Seeratan et al., 2020). In this context, planning skills deserve special attention, as they represent a cognitively demanding yet crucial stage of the inquiry process (Pedaste et al., 2015). Research planning requires the integration of subject-matter expertise with an understanding of the principles of scientific methodology and the ability to anticipate the potential outcomes and limitations of the proposed approach (Crujeiras-Pérez & Jiménez-Aleixandre, 2017; Strat et al., 2024). Inadequate planning skills can lead to methodological errors that negatively affect both the validity of the data collected and its subsequent interpretation (Bewersdorff et al., 2023; Kranz et al., 2023). In the study by Pedaste et al. (2021), planning skills are understood in a narrower sense as skills related to the design of a research procedure—that is, the transition from the formulation of a research problem to the empirical implementation of the investigation. This approach corresponds to the operationalization of planning skills in a concept that distinguishes several key categories of planning, such as the formulation of a research question or hypothesis, the identification and control of variables, the design of data collection methods and procedures, and the anticipation of possible experimental outcomes (Pedaste et al., 2021). Focusing on this phase allows for a more detailed analysis of how students design the structure of their research and what typical difficulties arise in this part of the inquiry process (Pedaste et al., 2021). Within inquiry cycle models (e.g., orientation – conceptualization – investigation – conclusion – discussion), planning serves as a key transition between problem formulation and the empirical conduct of research, and thus significantly determines the quality of the entire process (Pedaste et al., 2015). This model of sequential phases is a certain didactic simplification (Windschitl et al., 2008); in reality, the development of scientific knowledge is more iterative and pluralistic, involving ongoing evaluation and interpretation of the results obtained (see Harlen, 2021). However, it can provide a foundation for thinking about scientific inquiry as a response to a specific problem, which is a key element for further, more complex reflections on scientific epistemology.

2.3 Planning skills development

Despite the acknowledged importance of developing inquiry-related skills, many educational contexts continue to be dominated by transmissive teaching approaches focused on the reproduction of factual knowledge (Karlsen et al., 2025). Inquiry activities are often carried out on a limited basis or without sufficient methodological support (Strat et al., 2024). Furthermore, secondary analyses of international surveys suggest that the frequency of Inquiry-based learning activities alone does not necessarily lead to improved student performance in science literacy (cf. Jerrim et al., 2022, Oliver et al., 2021, Sjøberg,

2018). This discrepancy highlights the importance of the quality of implementation and the need for targeted scaffolding for the development of specific components of inquiry-related skills, rather than simply increasing the number of inquiry activities (Kang, 2022; Van Uum et al., 2017). As Riga et al. (2017) note, simply using the inquiry cycle may not be sufficient for the effective implementation of inquiry-based science education. One promising approach is the systematic use of inquiry-based learning tasks explicitly designed to develop planning skills (Schwichow et al., 2022). Tasks designed in this way should be based on a meaningful, context-based problem; they should require the formulation of a research question or hypothesis, the identification and operationalization of variables, and the design of a methodological approach (Nejedlý & Vojíř, 2024; Volkmann & Abell, 2003). It appears necessary to create a framework that will enable students to gradually take responsibility for the individual steps of the planning process and reflect on the limitations of the proposed research (Lombardi et al., 2018). Research also suggests that the impact of inquiry-based activities on planning skills represents a significant, but yet under-explored, aspect of inquiry-based education (Chen & Chen, 2025). The effectiveness of this approach depends not only on designing tasks considering the various stages of the inquiry process, but also on a deeper analysis of how students solve these tasks (Lazonder & Harmsen, 2016). It is becoming clear that, in order to understand the reasons behind students' success or failure in solving inquiry-based tasks, it is essential to focus on evaluating student performance at each stage of the inquiry process and to systematically identify the typical mistakes students make (Pedaste et al., 2021). This approach makes it possible to assess the level of students' initial planning skills, thereby providing a basis for targeted interventions that support the development of their scientific thinking.

3 Methodology

To answer the research questions, a mixed-methods study was conducted in 2024 at seven Czech elementary schools. The study combined quantitative analysis of the frequency with which students included selected aspects of inquiry design in their written plans with qualitative analysis of typical patterns in students' proposed methodological procedures.

3.1 Inquiry-based biology tasks

The study involved six inquiry-based tasks designed to encourage students to develop their own methodological approach to conducting inquiry. The biology tasks were piloted as part of the study by Nejedlý & Vojíř (2024).

Each inquiry task was designed according to a standard format. Each task always included:

1. Thematic motivational text;
2. instructions for students and the assigned research question;
3. an open space for students to propose an inquiry methodology;
4. opportunity for students to suggest the tools and materials to be used.

A uniform task structure was chosen to minimize the cognitive load associated with understanding the assignment, so that students could focus their attention on the planning process itself. Students worked with a given research question, and their task was to propose a methodological approach to investigating it, including research materials and tools. The assignment intentionally did not include any further specifics regarding the design of the methodological procedure.

The tasks represented two basic types of scientific research based on the research question: a) causal and b) descriptive. Each of these types was represented by three tasks.

Tasks containing a descriptive type of research problem:

- Observation of a slug – What are the length, width, and surface texture of a Spanish Slug?
- Observation of blooms – What are the colour, size, and number of petals Snowdrops and Spring Snowflakes blooms?
- Observation of fingers – What length are the fingers on the left and right hands of 14-years-old students?

These research questions were intended to guide students in developing a systematic approach to observing biological specimens and in designing an appropriate method for collecting and recording data. Students were asked to propose a comprehensive observation procedure that would allow them to answer the given question empirically.

Tasks involving a causal type of research problem:

- Experiment with yeast – How does temperature of the environment affect yeast activity?
- Experiment with a runner – How does running 800 meters affect a runner's body temperature?

- Experiment with tomatoes – How does enriching the soil with horse manure affect the weight of tomato fruit?

These research questions were intended to guide students in considering how to conduct systematic experiments in relation to the variables contained in the research question and in designing an appropriate method for collecting and recording data. Students were asked to propose a comprehensive experimental procedure that would enable them to answer the given question.

3.2 Assignment procedure and research sample

As part of the study, 15-year-old students gradually completed all six inquiry-based tasks during their regular biology classes. A total of 207 students initially participated in the study. Of these, 156 students were excluded from the analysis because they did not complete all six tasks. The final analytical sample therefore consisted of 51 students, each of whom completed all six inquiry-based tasks. Because every student solved each of the six tasks, the data have a repeated-measures structure: the same students contributed responses across all task contexts and both types of research problems. Before the study began, all teachers were trained in how to implement the inquiry-based tasks and received detailed methodological guidelines to ensure consistency in the study's implementation and adherence to the principle of not providing students with feedback on the tasks. To preserve mental capacity, students worked on only one task per class period. Biology classes were scheduled back-to-back to minimize external influences, and teachers did not provide direct feedback between tasks. Students worked on the tasks individually. The tasks were presented to them in a fixed order: Observation of a slug – Experiment with yeast – Observation of blooms – Experiment with a runner – Observation of fingers – Experiment with tomatoes. This order was chosen deliberately so that tasks involving descriptive and causal problems would alternate.

The study was conducted at seven public elementary schools. All schools followed the national curriculum, and science education was delivered in the format typical in the Czech Republic, namely as separate subjects of biology, chemistry, and physics. It was confirmed that the students at these schools had no prior educational experience with solving problems of this thematic nature and teachers at these schools do not use inquiry-based education in their lessons. Biology classes at all participating schools were taught by fully qualified teachers. The schools were selected with regard to sociodemographic diversity and the potential influences of the educational environment (Gay et al., 2019). The breakdown included the regions of Bohemia (5) and Moravia (2), specifically the Ústí nad Labem region (1), the Pardubice region (1), the Prague region (3), the South Moravia region (1), and the Central Bohemia region (1), as well as school sizes: up to 350 students (2), up to 700 students (3), and up to 1,000 students (2). The analysis included solutions from students who completed all 6 tasks. In total, solution sets from 51 students were further analysed.

3.3 Analysis procedure

The analysis of the students' solutions focused on evaluating the aspects of the inquiry procedure plans that the students proposed to answer the assigned research questions. Each solution was analysed using a tool based on categories formulated according to the EDAT framework (see Sirum & Humburg, 2011). This framework has already been tested in studies focused on assessing experimental design skills (e.g. Shanks et al., 2017). The aspects under consideration are also consistent with other definitions of research plan aspects in inquiry-based learning (see Fugarasti et al., 2019; McComas, 2014; National Research Council, 2013). The aspects evaluated in the students' solutions are listed in Table 1. Since it appears that students often find it difficult to identify the nature of the research problem when dealing with research tasks (Nejedlý & Vojíř, 2025), all aspects were evaluated for every task. Even in the case of tasks with a descriptive type of research problem, it was monitored whether students included irrelevant aspects, such as manipulated variables, in their solutions.

In the first phase of the analysis, closed coding was used. For each aspect, it was recorded whether it was **included** or **not included** in the student solution under review. Based on this coding, it was possible to determine the proportion of solutions in which students had taken the given aspect into account. To increase reliability, the coding was performed by two researchers. In the second phase, the analysis was expanded to include open coding focused on specific ways in which students addressed the assessed aspects of the methodological plan leading to the answer to the research question in individual tasks. To identify typical solution patterns, open codes were assigned to individual student solutions for each assessed aspect. Coding proceeded iteratively in a cyclical sequence, with solutions repeatedly revised and refined. The coding process continued until theoretical saturation was reached, at which point no new categories of meaning were identified. This process was carried out by two researchers, who continuously compared their coding and reached consensus in cases of disagreement.

Table 1: Aspects of inquiry design and characteristics of the evaluation

Aspects of Inquiry Design	*Characteristics of the evaluation*
inquiry approach	Whether the student proposed a procedure based on the systematic collection of empirical data through observation or experimentation. E.g. the student mentioned measuring a runner's temperature or observing differences in spring snowflake blooms.
manipulated variables	Whether the student identified a variable whose value they will intentionally change during the experiment. E.g., changing the ambient temperature when studying yeast activity, or adding manure to the soil when growing tomatoes.
observed variables	Whether the student has identified the variables or characteristics whose values will be recorded during the study. E.g., the length and width of a slug's body, the number of petals on blooms, or a runner's body temperature.
method of variable observation	Whether the student specified a specific method for measuring or observing the variables under study. E.g., measuring finger length with a ruler in centimetres, or measuring temperature with a thermometer before, during, and after a run.
constant conditions	Whether the student specified conditions that should be kept constant during the experiment. E.g., the same amount of yeast in the test tubes, the same amount of water, or the same measurement intervals.
sample characteristics	Whether the student described the basic characteristics of the objects or subjects on which the research will be conducted. E.g., the Spanish slug, snowdrop and snowflake blooms, 14-year-old students.
sample selection	Whether the student specified the method used to select the objects or subjects for the study. E.g., undamaged blooms, volunteers among 14-year-old students, or several specimens of slugs.
Repetition	Whether the student suggested repeating the measurement, observation, or experiment. E.g., The measurement was performed on five slugs, or the temperature was measured three times after the run.
data recording method	Whether the student described a method for systematically recording the collected data. E.g., entering values into a table or a graph.
tools and materials	Whether the student listed the instruments or supplies needed to conduct the research. E.g., ruler, magnifying glass, thermometer, stopwatch, test tubes, scale.
data analysis	Whether the student described how the collected data will be analysed or compared. For example, this could involve comparing values, calculating averages, identifying trends, or drawing conclusions based on the data.
uncertainty and error reduction	Whether the student considered potential sources of error and proposed methods to improve the reliability of the results. E.g., repeated measurements, a control sample, or verification of the results on multiple specimens.

3.4 Data processing and analysis

The data were descriptively analysed, standardized, and prepared for further analysis in MS Excel. In the analysis of data obtained through closed coding for the selected aspects, the percentage frequency of the inclusion of individual aspects in the entire set of solutions was first calculated based on the closed dichotomous coding. To further analyse the inclusion of procedural aspects according to the type of research problem, the occurrence of each aspect in tasks requiring experimental and observational approaches was calculated as a simple sum.

IBM SPSS Statistics software was used for the statistical analysis. As the Shapiro-Wilk test indicated significant deviations from normality for all items ($p < 0.01$), nonparametric tests were used in the analysis. Differences in the inclusion of individual aspects between students' solutions to tasks involving descriptive and causal problems were tested using the Wilcoxon signed-rank test. The use of the Wilcoxon signed-rank test and Cochran's Q test reflected the repeated-measures design of the study, as the same students provided responses to all six tasks. The Rosenthal correlation coefficient r was used to assess effect size. Differences in the inclusion of individual aspects in students' solutions across tasks were tested using Cochran's Q test, which allows for the comparison of multiple related binary values. In this study, the test examined whether a given aspect was included in the solutions to several tasks. Effect size was expressed using η^2. Effect size coefficients were interpreted according to Cohen (1988). Statistical significance was assessed at a p-value of < 0.05.

For data obtained through open coding, the most frequent combinations of student solutions were quantified. This identified typical patterns of methodological approaches used to answer the research questions in individual tasks. These solutions were described qualitatively, and representative examples were selected for them.

4 Results

4.1 Differences between task types in specific aspects

The vast majority of students chose an inquiry-based approach in their inquiry plans to answer the assigned research questions. However, when considering the individual aspects examined, significant differences were found in the inquiry plans; see Table 2. In the vast majority of cases, students included the tools and/or materials they plan to use for their inquiry and considered the variables whose values they will monitor.

However, students specified the specific method for tracking these variables or provided a detailed description of the research sample on which they would conduct their study in only about half of their solutions. Other aspects were rarely considered. Students included methods for evaluating the collected data only very sporadically, and after describing the setup of the observation or experiment, they typically provided general statements such as: "I will write everything down and answer the research question." In their plans, students practically did not address solutions to potential limitations of the research or procedures for reducing errors. This aspect was addressed only in a handful of cases where students chose to use a control sample, repeated control measurements, verification of the procedure's functionality on a different sample, measurements using different devices and comparison of findings, or a review by another researcher.

Table 2: Consideration of individual aspects in student inquiry plans and statistical significance of differences between task solutions. Statistically significant p-values are marked in bold. Rosenthal r and eta squared values indicating medium effect sizes are marked in italics, and values indicating large effect sizes are marked in bold.

	proportions in solutions			difference according research problem type			difference among particular tasks		
	All tasks	Descriptive problems	Causal problems	Z	p	r	Q	p	η^2
inquiry approach	99%	98%	99%	−1.41	0.16	0.20	12.22	**0.03**	0.05
manipulated variables	44%	1%	88%	−6.51	< **0.01**	**0.91**	212.49	< **0.01**	**0.83**
observed variables	79%	84%	75%	−2.45	**0.01**	*0.34*	41.51	< **0.01**	**0.16**
method of variable observation	53%	58%	47%	−2.44	**0.02**	*0.34*	23.58	< **0.01**	*0.09*
constant conditions	14%	8%	20%	−3.48	< **0.01**	*0.49*	32.57	< **0.01**	*0.13*
sample characteristics	53%	53%	54%	−0.17	0.87	0.02	35.55	< **0.01**	*0.14*
sample selection	14%	27%	2%	−4.37	< **0.01**	**0.61**	62.86	< **0.01**	**0.25**
Repetition	26%	40%	12%	−4.23	< **0.01**	**0.59**	55.00	< **0.01**	**0.22**
data recording method	24%	29%	19%	−2.58	**0.01**	*0.36*	9.52	0.09	0.04
tools and materials	98%	97%	98%	−0.58	0.56	0.08	12.04	**0.03**	0.05
data analysis	4%	1%	7%	−2.71	**0.01**	*0.38*	1.95	0.06	0.04
uncertainty and error reduction	3%	2%	4%	−1.00	0.32	0.14	6.95	0.22	0.03

Several statistically significant differences were found in how students considered individual aspects in tasks requiring an experimental approach and an observational approach without manipulating variables. Statistically significant differences were absent only for aspects that students considered either almost universally or almost never; see Table 2. Furthermore, no statistically significant difference was found in the case of the characteristics of the research sample, which slightly exceeded half of the solutions in both types of approaches. The most pronounced difference was found in the inclusion of manipulated variables. This difference stems from the nature of the research questions posed; in the case of descriptive research questions, students failed to correctly mention variable manipulation in the vast majority of cases. Conversely, in most plans for answering research questions involving a causal type of research problem, students correctly reported variable manipulation. Other significant differences of large effect size were found regarding research sample selection and replication of the procedure. In both cases, students more frequently mentioned these aspects in procedures related to descriptive problems. Here, students more frequently cited the need to select a representative sample from a larger population that, for example, would not be biased. They mentioned selecting a larger number of subjects for observation. In contrast, when planning an experiment to address a causal problem, sample selection was practically non-existent, and only in a small number of cases did students mention conducting experiments on a larger number of samples or repeating the procedure.

4.2 Differences between individual tasks in specific aspects

A more detailed analysis further revealed differences between individual tasks. In the case of tasks with a descriptive problem type, the frequency of mentioning the selection of the research sample varied considerably (see Fig. 1). This aspect was very rarely mentioned in the observation of the slug (8%), while it was mentioned more frequently in the observation of blooms (35%) and fingers (37%). The greatest differences in these tasks were found in constant conditions. These were most frequently considered in the observation of fingers (20%), while in other tasks focused on description, this aspect did not appear at all or only rarely (4%).

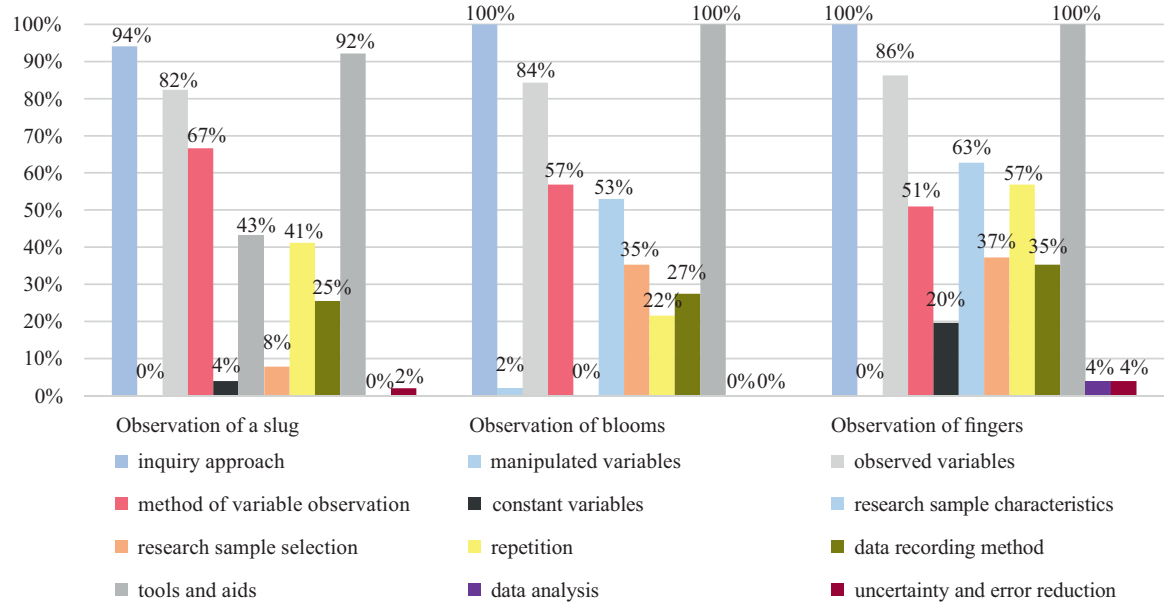

Fig. 1: The proportion of solutions containing the assessed aspects in tasks with a descriptive problem type

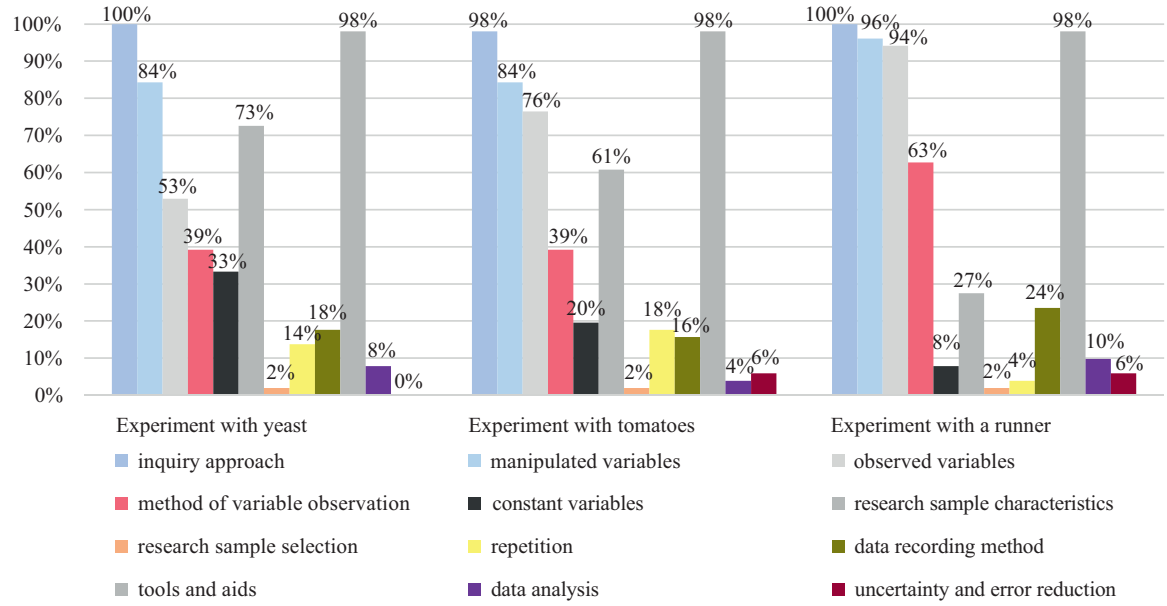

Fig. 2: The proportion of solutions containing the assessed aspects in tasks with a causal problem type

The inclusion of the constant conditions aspect was strongly influenced by the specific topic of the problem being addressed in tasks focused on causal relationships (see Fig. 2). While it was considered in one-third of the cases in the yeast experiment, it was considered in only 8% of the cases in the runner experiment. Furthermore, significant differences were found in these tasks regarding the reporting of observed variables. While the observed variable was reported in 94% of solutions for the runner experiment, it was 76% for the tomato experiment and only 53% for the yeast experiment. Conversely, in the runner task, students more frequently specified the method of monitoring the variable (63%) compared to the

yeast and tomato experiments (both 39%). A significant divergence in experimental design plans was also found in the characterization of the research sample. The highest representation of this aspect was recorded in the yeast experiment (73%), while it was lowest in the runner experiment (27%). A similar trend of low representation in the runner experiment was also found for the aspect of repetition, which was included in only 4% of the solutions.

4.3 Typical solutions based on the considered aspects

4.3.1 Observation of a slug

In the case of the task focused on observing the Spanish slug, 17 types of student solutions were identified based on the aspects of the inquiry design that were considered. The most common type of solution (23% of student solutions) was a simple proposal for an observation procedure focused on a single Spanish slug.

> *After the rain, we'll find a Spanish slug. We'll use a ruler to measure its length and width. We'll use a magnifying glass to examine the slug's structure. We'll record our findings. Then we'll release the slug back into the wild.*
>
> *Materials and tools: ruler, Spanish slug, magnifying glass.*

In this solution, the students adopted an inquiry-based approach and identified the key variables of interest, specifically the length, width, and surface structure of the body. They also listed the tools and aids they used, primarily a ruler for measuring dimensions and a magnifying glass for closer observation of the body's surface. The procedure also included recording the information gathered, which indicates at least a basic understanding of data documentation.

From a methodological standpoint, however, this approach remains rather limited. The students did not describe the characteristics of the research sample, as the observation involved only a single specimen. Furthermore, there was no repetition of measurements, which would have allowed for data comparison and increased the reliability of the results. The solution also did not include a detailed specification of how the variables would be monitored, such as how body length would be measured given the varying dimensions of different body parts or considering the slug's movement, nor how the observed traits would be systematically recorded or analysed. Overall, this is a simple, intuitive observation design that allows for the collection of basic descriptive data but does not provide a sufficient methodological framework for more systematic biological investigation.

The second type of common student solution, which appeared with the same frequency (23% of student solutions), was significantly more sophisticated in terms of inquiry design.

> *We'll collect Spanish slugs in a bowl. I'll prepare a mat, paper, or newspaper and some string. We'll transfer one slug onto the mat, paper, or newspaper. We'll observe its movements as it stretches and contracts. I'll take the string and measure the length and width of one slug. We record the measured values. We perform the measurements on two more specimens. We examine the structure of the slug with a magnifying glass. We record our observations in the log. I make sure not to mix up the millipedes. I place them in different bowls. We evaluate the data. We answer the research question.*
>
> *Materials and tools: ruler, spiral notebook, 3× Spanish slugs, paper, pencil, 3× bowls, mat, paper/newspaper, magnifying glass, string.*

In this case, the students defined a larger research sample and incorporated repeated measurements. They also described how individual specimens would be handled and separated and specified the use of string to measure body dimensions despite the slug's movement and changing body shape. Compared with the first solution, this design also includes more explicit data recording and subsequent evaluation. However, it still lacks a more systematic approach to data analysis.

4.3.2 Observation of blooms

In the task focused on observing bloom parts, 20 types of student solutions were identified based on the aspects of the inquiry design that were considered. The most common type of solution (18%) involved identifying the key variables under observation, namely the colour, size, and number of petals in the blooms of two plant species.

> *Count the petals on the blooms and record the number. Measure the size of the blooms with a ruler and record the measurement. Identify the colour. Pay close attention to all the results and answer the question.*
>
> *Materials and tools: blooms from the plants listed, a magnifying glass, and a ruler.*

The students proposed methods for monitoring variables, such as measuring bloom size with a ruler and counting the number of petals. However, this proposal remained very general, and the students did not specify, for example, which parts they would measure. The solution also listed the tools to be used—a ruler and a magnifying glass. The students stated that they would record the results but did not specify how they would keep records. This type of solution did not include repeating the observations on a larger number of specimens or a systematic description of data analysis, such as how to compare results among different plant species. From a methodological perspective, this is a relatively functional proposal for simple observation, but with limited depth in the inquiry design.

The second most common type of solution (16%) also considered sample selection, emphasizing that the blooms should be whole and undamaged. These solutions included observing a larger number of blooms and comparing them, which suggests an understanding of how appropriate object selection can affect the quality of results.

> *I will prepare both types of plants—avoiding damaged or incomplete blooms. I will examine the blooms and first assess their colour. I will pluck one petal and measure its size with a ruler. I'll record the colour and size in my notebook. I'll count the number of petals on each plant. I'll record the results. I'll evaluate and compare the results. I'll answer the research question.*
>
> *Materials and tools: two snowdrop blooms, two snowflake blooms, a pencil, a notebook, a ruler.*

4.3.3 Observation of fingers

In the task focused on observing fingers, 27 types of student solutions were identified based on the aspects of the inquiry design that were considered. The most common type of solution (16%) involved only the basic identification of the research sample—namely, 14-year-old students—and a proposal to measure the length of their fingers using a ruler.

> *I'll bring in some 14-year-old students and ask them to help with the research. I'll measure their fingers from all sides. Done.*
>
> *Materials and tools: ruler, students.*

The students did not specify the exact method of measurement or the point from which the finger length would be measured. Their solution did not include repeated measurements, systematic data recording, or subsequent data analysis.

The second most common type of solution (14%), on the other hand, was significantly more complex.

> *I will ask the volunteers to undergo a measurement. I will use a tape measure or ruler to measure the length of their finger from the centre of the joint to the tip. I will take measurements on both the right and left hands. I will record the data in centimetres. I will repeat this process for all 14-year-olds. I will compare and analyse the data. I will answer the research question.*
>
> *Materials and tools: ruler, volunteers, paper, pencil.*

Unlike the first solution, this response specified the measurement method, defining finger length as the distance from the joint to the fingertip. It also identified a sample of 14-year-old volunteers, included repeated measurements across multiple individuals, and involved both data recording and subsequent comparison, representing a basic form of data analysis.

4.3.4 Experiment with yeast

In the task focused on the yeast experiment, 22 types of student solutions were identified based on the aspects of inquiry design that were considered. The most common type of solution (16%) was a relatively complex proposal for an experimental procedure.

> *I'll prepare a mixture of yeast, water, and sugar in two test tubes. We want the same amount of mixture in both. I'll fill the beakers with cold and warm water and grab a stopwatch. I'll label the beakers 'T' for warm water and 'S' for cold water. I place one test tube in each beaker. I record what happens at each minute. I record the readings after two minutes and after ten minutes.*
>
> *Materials and tools: sugar, 2 test tubes, water, a stopwatch, paper, a pencil, warm water.*

The students clearly identified the manipulated variable (ambient temperature). The procedure also included the method of monitoring the variable: observing changes over time and recording them at regular intervals. They also considered constant conditions, such as using the same amount of yeast, water, and sugar in each test tube, and included the necessary materials and a plan for data recording.

The second most common type of solution (14%) was significantly simpler. The students did not specify the dependent variable or how it would be measured. The solution also lacked data recording and subsequent analysis.

> *Label test tubes one through four. Place yeast and water in each test tube. Place the first and second test tubes in a cool environment and the third and fourth test tubes in a warm environment. The experiment will last ten minutes.*
>
> *Materials and tools: four test tubes.*

4.3.5 Experiment with tomatoes

In the task focused on the tomato experiment, 21 types of student solutions were identified based on the aspects of the inquiry design that were considered. However, only a single type of solution significantly outnumbered the others in this case. This most common solution (14%) included the identification of the manipulated variable—adding manure to the soil—and the observed variable, which was the weight of the tomato fruits.

> *We will plant the tomato seedlings. We will label the seedlings A through D. We will apply fertilizer to three of them. After the tomatoes have grown, we will weigh the fruits from each plant. I will record the weights, analyse the data, and answer the research question.*
>
> *Materials and tools: manure, 4 tomato seedlings, scale, pen, paper, water.*

The students also described how to measure using a scale and planned to record and analyse the data. The experimental design included a control group, since only some of the plants were fertilized. However, it did not explicitly specify constant conditions, such as equal amounts of water or light, or include replication with a larger number of plants.

4.3.6 Experiment with a runner

In the task focused on the runner experiment, 18 types of student solutions were identified based on the aspects of inquiry design that were considered. The most common type of solution (25%) was a simple experiment in which students identified the manipulated variable, namely an 800 meter run, and the observed variable, namely the runner's body temperature.

> *We will measure the runner's temperature. We will have the runner run 800 meters. Once they finish, we will measure their temperature with a thermometer. We will compare it to their original temperature before running.*
>
> *Materials and tools: thermometer, measuring tape.*

The procedure involved measuring body temperature before, during, and after the run and then comparing the results. However, the study did not include repeated measurements, the selection of a research sample, or the control of other factors that could influence body temperature.

The second most common type of solution (18%) expanded the design by including repeated temperature measurements at several intervals after the run.

> *First, we'll take the runner's temperature twice while he's at rest, then he'll run 800 meters. When he returns, we'll take his temperature right away. We'll take his temperature again after one minute, and again after two minutes. We'll compare the temperatures we've measured.*
>
> *Materials and tools: a thermometer, a place where you can run 800 meters.*

4.3.7 Summary students' inquiry proposal design

An analysis of typical solutions across individual tasks reveals a relatively consistent pattern in how students approached the design of a research procedure. In most cases, students were able to identify the basic research approach and correctly identified the variables under study and the corresponding instruments or tools needed to measure or observe them. These elements appeared in both descriptive tasks (e.g., observing a slug, blooms, or fingers) and experimental tasks (e.g., yeast, tomatoes, or a runner).

Students also frequently mentioned basic data recording, indicating a fundamental understanding of the need to document the results obtained.

In contrast, more complex aspects of inquiry design appeared significantly less frequently in their solutions. Only a portion of the students included a description of the research sample or the repetition of the procedure, which are important for increasing the reliability of the data obtained. Even less frequently did students explicitly mention constant experimental conditions or methods for controlling them. Another significant shortcoming was the absence of a more systematic data analysis plan that would allow for a structured comparison of the results obtained and the formulation of conclusions. Only in very rare cases did the solutions take measurement uncertainty into account or propose procedures for reducing measurement errors.

Overall, the results suggest that students are relatively proficient at identifying variables and the practical steps involved in data collection, but they consider methodological principles that ensure the reliability and interpretability of results significantly less often. This pattern emerged in both descriptive and causal inquiry-based tasks, suggesting that students' difficulties are related more to the more complex aspects of research planning than to the type of research question itself.

5 Discussion

The results show that, in most cases, 15-year-old Czech students correctly recognized that answering the research question requires an empirical inquiry approach; this is also confirmed by a study conducted by Teig (2024). The nearly universal choice of observation or experimentation may indicate that, within the context of these written tasks, students were able to associate the research questions with an empirical approach; this connection is also confirmed by e.g. Pedaste et al. (2015) and van Uum et al. (2016). This result is particularly evident in the distinction between descriptive and causal tasks: while students generally did not identify the manipulated variables in descriptive problems, they often correctly identified them in causal tasks. These results are consistent with the findings of a qualitative study by Arnold et al. (2014) regarding issues in students' procedural understanding. A possible cause, as noted by Schwichow et al. (2022), is insufficient metacognitive knowledge of the relationship between manipulative and observed variables. Tairab (2015) notes and confirms that students can, to some extent, distinguish between when it is appropriate to use observation and when to use experimentation; however, he also points out that guided inquiry – focused primarily on verifying and illustrating phenomena – predominates in the classroom, leading to its dominance in the learning context and limiting opportunities for authentic scientific inquiry. This finding corresponds with the fact that understanding remains largely at the level of basic procedure selection and is less reflected in its methodological sophistication (see Kranz et al., 2023).

The analyses showed that the students were relatively proficient, particularly in those aspects of the design that are directly related to specific activities, namely identifying the variables of interest and listing the instruments and tools (cf. Arnold et al., 2014). Typical solutions were often based on a brief description of measurements or observations and their recording; more complex proposals involving multiple measurements, sampling, or basic data analysis were less common. This pattern can be linked not only to the cognitive demands of individual aspects of the inquiry design (cf. Schwichow et al., 2022), but also with the nature of science education in schools, where students are often guided more toward reproducing predetermined "experiments" and demonstrations than toward independently thinking through the structure of inquiry, as confirmed by, e.g., Abrahams and Millar (2008) and Janštová and Pavlasová (2019). In other words, students have a fairly good idea of what they will do, but they rarely explicitly consider why the procedure should be designed this way and what conditions it must meet in order to lead to reliable conclusions (see Zimmerman, 2007). The acquisition of these patterns is apparently also influenced by the nature of the investigative tasks presented in the teaching materials. For example, in chemistry textbooks, laboratory activities are typically described only procedurally, as a list of mechanical steps accompanied by a corresponding list of equipment. However, aspects of methodological planning—such as linking the procedure to the research question, selecting and specifying samples, evaluating results, or eliminating errors and uncertainties—are not presented in textbooks (see Vojíř, 2021).

Even in the students' plans for methodological approaches, constant conditions, repetition of procedures, data analysis, and—in particular—dealing with uncertainty were taken into account significantly less often; this pattern is consistent with earlier research, which identifies precisely these components as problematic in the acquisition of inquiry-related skills (cf. Arnold et al., 2014; Schwichow et al., 2022). It is precisely here that shortcomings in epistemic knowledge and its application in inquiry design are likely to become apparent (see Sandoval, 2005). While identifying a variable or a tool is relatively straightforward, taking into account variability, sources of error, or how to interpret results requires a deeper understanding of the nature of science of scientific inquiry (see Zetterqvist & Bach, 2023; Zimmerman, 2007). The

very low prevalence of procedures for reducing uncertainty therefore suggests that students often view measurement as a means of obtaining a single "correct answer", rather than as an error-prone process that requires verification and repetition, which is consistent with the findings reported Kok (2022).

The differences between individual tasks further demonstrate that, in addition to the type of research problem, the context of the specific task also plays a significant role, as noted Brown (2019). For example, in the task involving the runner, students more frequently specified the method of monitoring the variable than in other experimental tasks, likely because body temperature is a familiar and well-established concept in school settings. In contrast, in the yeast task, the monitored variable was specified less frequently, even though it involved a methodologically rich experimental situation. This suggests that students find it easier to work with variables that are commonly used in school instruction and have established terminology, while they face greater difficulties when they need to independently specify a variable in a less standard context, a finding confirmed by other studies (e.g. Shaffer et al., 2016; Schäfer et al., 2024). It is therefore not merely a matter of identifying variables as such, but also of the ability to translate a specific biological phenomenon into explicit research language (see Hamnell-Pamment, 2024).

It is also interesting to note that sample selection and repetition of the procedure occurred more frequently in descriptive tasks than in causal tasks. This result suggests that students associated the need for multiple examples more with comparing objects than with verifying the reliability of experimental results, which is consistent with the findings reported Klahr and Chen (2011). When it came to blooms and fingers, they more often considered multiple instances, whereas for experimental tasks, they often settled for the idea of a single trial. Thus, in their written plans, students appeared to apply repeatability and representativeness more as task-specific considerations than as general principles of inquiry design (see Arnold et al., 2014).

5.1 Implications

Based on the research conducted, it appears that instruction should not be limited solely to incorporating inquiry tasks (cf. Lazonder & Harmsen, 2016), but should more explicitly support the individual components of inquiry planning. Although students in most cases correctly recognized that answering a research question requires an empirical approach, their proposals often remained at the level of immediate procedural steps. This suggests that the experience of an inquiry task alone may not automatically lead to the development of more methodologically sophisticated thinking (see Heindl, 2019; Jerrim et al., 2022). It therefore seems particularly important to focus instruction on those aspects of planning that students spontaneously considered only to a limited extent, particularly the control of conditions, repetition of procedures, data analysis, and dealing with uncertainty. These aspects likely require targeted scaffolding (see van Uum et al., 2017), which will help students move beyond merely describing "what they will do" toward thinking through why a given procedure is methodologically sound and under what conditions it can lead to reliable conclusions.

This need is further underscored by the importance of explicitly mapping out a inquiry plan. The results suggest that students are relatively proficient at identifying variables and tools, but they less frequently consider the relationships between them, the methods for analysing the data, or the limitations of the proposed procedure. According to Renkl (2014), it can be pedagogically beneficial to make the structure of an inquiry design accessible through model solutions, joint analysis of examples, or a comparative examination of methodologically stronger and weaker designs. Such model solutions can serve not only as a support for students but also as a support for teachers in focusing attention on the key features of different types of tasks. In this sense, they can function as scaffolding tools that structure both student learning and pedagogical decision-making regarding which aspects of planning need to be emphasized in a specific type of task (see van Uum et al., 2017). It seems that it might be useful to break down inquiry planning into explicit steps and provide targeted questions or supporting tools for each of them. Such an approach could help ensure that the methodologically and epistemologically more demanding components of inquiry design become part of students' everyday thinking gradually, as noted by, for example, Sandoval (2005) and Zimmerman (2007), rather than as implicit expectations.

The results show that the quality of students' solutions is not determined solely by their general skill level but is significantly influenced by the specific thematic context of the task. Differences between individual tasks suggest that students find it easier to work with variables and procedures that are commonly used in school instruction and linguistically established, whereas in less familiar contexts they have greater difficulty with their explicit formulation and operationalization (see Shaffer et al., 2016). It follows that the development of planning skills should not be tied to individual, isolated activities or specific subject areas, but should be systematically supported across various science topics. What is important here is not only the repetition of inquiry tasks, but also the variability of contexts, which allows students to transfer planning strategies between different types of problems and gradually abstract them

from specific content. The study's findings thus support the need to design inquiry tasks not only in terms of their thematic content but also in terms of the specific planning skills and epistemic understanding they are intended to develop in students.

5.2 Limitations of the study

When interpreting the results, several limitations of the study must be considered. The research was conducted using a set of six research-oriented biology tasks, which limits the generalizability of the findings to other areas of science and types of tasks. Differences between tasks may thus have been influenced not only by the type of research problem but also by the specific content and the students' familiarity with the given topic. Although the order was chosen deliberately to alternate descriptive and causal research problems, possible order effects cannot be ruled out. Students' responses to later tasks may have been influenced by their experience with previous tasks.

Furthermore, the study analysed written proposals for methodological procedures, not the actual conduct of the research. The results therefore reflect the students' ability to explicitly articulate a research plan, not necessarily all aspects they would consider in practical work. While the open-ended nature of the assignment, without detailed guidance, it may have led to the underestimation of certain aspects of the research design that students did not consider necessary to explicitly mention.

Another limitation is the size of the analysed dataset, which included only 51 complete sets of solutions; this may limit the generalizability of the results and suggests possible sample bias. These limitations do not diminish the significance of the findings, but they do constrain their interpretation and highlight the need for further research, particularly linking the analysis of proposals to actual research activities and verifying the results on larger and more diverse student samples.

6 Conclusion

This study contributes to our understanding of how 15-year-old lower-secondary school students design inquiry procedures when solving inquiry-based biology tasks. The results show that, in most cases, students correctly recognize that answering the given research question requires an empirical inquiry approach, and at a basic level, they can distinguish between descriptive and causal types of problems. However, it appears that their proposals remain largely at the level of immediate procedural steps. Most frequently, they included the identification of observed variables and the specification of instruments or tools, while components requiring deeper methodological and epistemic understanding—particularly the control of conditions, repetition of the procedure, data analysis, and dealing with uncertainty—appeared significantly less often. Furthermore, the differences observed between individual tasks suggest that students' performance is determined not only by the type of research question but also by the specific thematic context and the extent to which the given variables or procedures are familiar to students from school and part of their vocabulary.

These results suggest that the development of planning skills cannot be reduced to merely incorporating inquiry-based activities but rather requires targeted and explicit support focused on the structure of the research design. From a didactic perspective, it therefore seems important to pay greater attention to those aspects of planning that students do not spontaneously include, and to support their development across different types of tasks and subject areas. Further research should therefore examine how planning skills change over time, how they develop gradually through inquiry-based tasks, how they are influenced by various forms of scaffolding support, and to what extent they transfer across different science contexts.

Declaration of interest statement

No conflict of interest to declare.

Ethics statement

We confirm that the study was carried out in accordance with the ethical principles outlined in the Belmont Report and complied with all relevant ethical, legal, and human subject protection standards applicable to this type of research. Informed consent was obtained from all participating schools prior to the inclusion of students in the study. All personal data were fully anonymised, and researchers had access only to coded identifiers that did not allow for the identification of individual participants or schools. During manuscript preparation, the authors used ChatGPT solely for AI-assisted copy editing to improve readability, style, grammar, spelling, punctuation, and tone. It was not used for content creation, research ideas, data analysis, interpretation, or any substantive scholarly contribution.

Funding

This publication was supported by Institutional Support for Longterm Development of Research Organizations – Cooperatio SOC/Subject Specific Education Research – Charles University, Faculty of Education.

References

Abrahams, I., & Millar, R. (2008). Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education*, *30*(14), 1945–1969. https://doi.org/10.1080/09500690701749305

Arnold, J. C., Kremer, K., & Mayer, J. (2014). Understanding students' experiments—what kind of support do they need in inquiry tasks? *International Journal of Science Education*, *36*(16), 2719–2749. https://doi.org/10.1080/09500693.2014.930209

Baltikian, M., Kärkkäinen, S., & Kukkonen, J. (2024). Assessment of scientific literacy levels among secondary school students in Lebanon: exploring gender-based differences. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(3), em2407. https://doi.org/10.29333/ejmste/14279

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: a comparative study with human raters. *Computers and Education: Artificial Intelligence*, *5*, 100177. https://doi.org/10.1016/j.caeai.2023.100177

Brown, J. P. (2019). Real-World Task Context: Meanings and Roles. In G. A. Stillman & J. P. Brown (Eds.), *Lines of inquiry in mathematical modelling research in education* (pp. 53–81). Springer International Publishing. https://doi.org/10.1007/978-3-030-14931-4_4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Routledge. https://doi.org/10.4324/9780203771587

Crujeiras-Pérez, B., & Jiménez-Aleixandre, M. P. (2017). High school students' engagement in planning investigations: findings from a longitudinal study in Spain [10.1039/C6RP00185H]. *Chemistry Education Research and Practice*, *18*(1), 99–112. https://doi.org/10.1039/C6RP00185H

Čipková, E., Šmida, D., & Pecníková, K. (2025). Evaluation of the level of selected inquiry skills among grammar school students. *Science & Education*, *34*(5), 3727–3749. https://doi.org/10.1007/s11191-024-00602-3

Elhai, J. (2023). Science literacy: a more fundamental meaning. *Journal of Microbiology & Biology Education*, *24*(1), e00212-00222. https://doi.org/10.1128/jmbe.00212-22

Fugarasti, H., Ramli, M., & Muzzazinah. (2019). Undergraduate students' science process skills: a systematic review. In N. Y. Indryanti, M. Ramli, & F. Nurhasanah (Eds.), *AIP Conference proceedings: Vol. 2194 (1). 2nd International conference on Science, Mathematics, Environment, and Education.* AIP Publishing. https://doi.org/10.1063/1.5139762

Gay, L. R., Mills, G. E., & Airasian, P. W. (2019). *Educational research: competencies for analysis and applications.* Pearson.

Hamnell-Pamment, Y. (2024). The role of scientific language use and achievement level in student sensemaking. *International Journal of Science and Mathematics Education*, *22*(4), 737–763. https://doi.org/10.1007/s10763-023-10405-7

Harlen, W. (2021). *The case for inquiry-based science education (IBSE).* AIP. https://www.interacademies.org/publication/case-inquiry-based-science-education-ibse

Heindl, M. (2019). Inquiry-based learning and the pre-requisite for its use in science at school: a meta-analysis. *Journal of Pedagogical Research*, *3(2)*, 52–61. https://doi.org/10.33902/JPR.2019254160

Chen, F., & Chen, G. (2025). Learning analytics in inquiry-based learning: a systematic review. Educational technology research and development, *73*(4), 2131–2161. https://doi.org/10.1007/s11423-025-10507-9

Janštová, V., & Pavlasová, L. (2019). Inquiry vs. cookbooks in practical teaching biology viewed by teachers. In M. Rusek & K. Vojíř (Eds.), *Project-based education and other activating strategies in science education XVI* (pp. 30–36). https://page.pedf.cuni.cz/pbe/files/2019/07/sbornikPBE2018_wos.pdf

Jerrim, J., Oliver, M., & Sims, S. (2022). The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England. *Learning and Instruction*, *80*, 101310. https://doi.org/10.1016/j.learninstruc.2020.101310

Ješková, Z., Lukáč, S., Šnajder, Ľ., Guniš, J., Klein, D., & Kireš, M. (2022). Active learning in STEM education with regard to the development of inquiry skills. *Education Sciences*, *12*(10), 686. https://doi.org/10.3390/educsci12100686

Kang, J. (2022). Interrelationship between inquiry-based learning and instructional quality in predicting science literacy. *Research in Science Education*, *52*(1), 339–355. https://doi.org/10.1007/s11165-020-09946-6

Karlsen, S., Kersting, M., Ødegaard, M., Olufsen, M., Lill Suhr, M., & Kjærnsli, M. (2025). A comparative study of teachers' conceptualisations and enactment of inquiry-based science education. *International Journal of Science Education*, 1–31. https://doi.org/10.1080/09500693.2025.2547411

Klahr, D., & Chen, Z. (2011). Finding one's place in transfer space. *Child Development Perspectives*, *5*(3), 196–204. https://doi.org/10.1111/j.1750-8606.2011.00171.x

Kok, K. (2022). *Certain about uncertainty* [Dissertation, Humboldt-Universität zu Berlin]. https://doi.org/10.18452/24782

Kranz, J., Baur, A., & Möller, A. (2023). Learners' challenges in understanding and performing experiments: a systematic review of the literature. *Studies in Science Education*, *59*(2), 321–367. https://doi.org/10.1080/03057267.2022.2138151

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Review of Educational Research*, *86*(3), 681–718. https://doi.org/10.3102/0034654315627366

Lombardi, D., Bailey, J. M., Bickel, E. S., & Burrell, S. (2018). Scaffolding scientific thinking: students' evaluations and judgments during Earth science knowledge construction. *Contemporary Educational Psychology*, *54*, 184–198. https://doi.org/10.1016/j.cedpsych.2018.06.008

McComas, W. F. (2014). Science process skills. In W. F. McComas (Ed.), *The language of science education: An expanded glossary of key terms and concepts in science teaching and learning* (pp. 89–89). SensePublishers. https://doi.org/10.1007/978-94-6209-497-0_79

Mocanu, M., Bibiri, A.-D., Rusu, V. D., Moroşanu, A., & Bejan, I. G. (2025). Enhancing civic engagement with science: a comparative approach across European regions. *Scientometrics*, *130*(1), 447–468. https://doi.org/10.1007/s11192-024-05198-7

NASEM. (2016). *Science Literacy: concepts, contexts, and consequences.* The National Academies Press. https://doi.org/10.17226/23595

National Research Council. (2012). *A Framework for K-12 science education: practices, crosscutting concepts, and core ideas.* The National Academies Press. https://doi.org/10.17226/13165

National Research Council. (2013). *Next generation science standards: for states, by states.* The National Academies Press. https://doi.org/10.17226/18290

Nejedlý, A., & Vojíř, K. (2022). How do students formulate a research question and conclusions in science research? In M. Rusek & M. Tóthová (Eds.), *Project-based education and other student-activation strategies and issues in science education XIX.* (pp. 29–38). Charles University, Faculty of Education. https://pages.pedf.cuni.cz/pbe/files/2022/10/ProceedingsPBE2021_final.pdf

Nejedlý, A., & Vojíř, K. (2024). How to construct inquiry tasks: a methodological framework and task piloting. In D. Koperová & M. Rusek (Eds.), *Project-based education and other activating strategies in science education XXI.* (pp. 125–135). Charles University, Faculty of Education. https://pages.pedf.cuni.cz/pbe/files/2024/09/PBE2023_final.pdf

Nejedlý, A., & Vojíř, K. (2025). The role of inquiry-based learning: development of analytical skills in orientation and conceptualization by science inquiry at lower secondary school. In G. L. Chova, M. G. Ch., & L. J. (Eds.), *17th International conference on education and new learning technologies* (Vol. 17, pp. 7494–7501). IATED Academy. https://doi.org/10.21125/edulearn.2025.1845

OECD. (2023a). *PISA 2022 results (volume I): The state of learning and equity in education.* OECD Publishing. https://doi.org/10.1787/53f23881-en

OECD. (2023b). *PISA 2025 Science framework draft.* https://pisa-framework.oecd.org/science-2025/assets/docs/PISA_2025_Science_Framework.pdf

OECD. (2025). *OECD science, technology and innovation outlook 2025: driving change in a shifting landscape.* https://doi.org/10.1787/5fe57b90-en

Oliver, M., McConney, A., & Woods-McConney, A. (2021). The efficacy of inquiry-based instruction in science: a comparative analysis of six countries using PISA 2015. *Research in Science Educcation, 51* (Suppl 2), 595–616. https://doi.org/10.1007/s11165-019-09901-0

Osborne, J., & Allchin, D. (2024). Science literacy in the twenty-first century: informed trust and the competent outsider. *International Journal of Science Education*, *47*(15–16), 1–22. https://doi.org/10.1080/09500693.2024.2331980

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections.* King's College London. https://www.nuffieldfoundation.org/wp-content/uploads/2019/12/Sci_Ed_in_Europe_Report_Final1.pdf

Pedaste, M., Baucal, A., & Reisenbuk, E. (2021). Towards a science inquiry test in primary education: development of items and scales. *International Journal of Stem Education*, *8*(1), 19. https://doi.org/10.1186/s40594-021-00278-z

Pedaste, M., Maeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, *38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Riga, F., Winterbottom, M., Harris, E., & Newby, L. (2017). Inquiry-based science education. In K. S. Taber & B. Akpan (Eds.), *Science education: An international course companion* (pp. 247–261). SensePublishers. https://doi.org/10.1007/978-94-6300-749-8_19

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, *89*(4), 634–656. https://doi.org/10.1002/sce.20065

Seeratan, K. L., McElhaney, K. W., Mislevy, J., McGhee Jr, R., Conger, D., & Long, M. C. (2020). Measuring students' ability to engage in scientific inquiry: A new instrument to assess data analysis, explanation, and argumentation. *Educational Assessment*, *25*(2), 112–135. https://doi.org/10.1080/10627197.2020.1756253

Shaffer, J. F., Dang, J. V., Lee, A. K., Dacanay, S. J., Alam, U., Wong, H. Y., Richards, G. J., Kadandale, P., & Sato, B. K. (2016). A familiar(ity) problem: assessing the impact of prerequisites and content familiarity on student learning. *PLoS One*, *11*(1), e0148051. https://doi.org/10.1371/journal.pone.0148051

Shanks, R. A., Robertson, C. L., Haygood, C. S., Herdliksa, A. M., Herdliska, H. R., & Lloyd, S. A. (2017). Measuring and advancing experimental design ability in an Introductory course without altering existing lab curriculum. *Journal of microbiology & biology education*, *18*(1), 1–8. https://doi.org/10.1128/jmbe.v18i1.1194

Schäfer, J., Reuter, T., Karbach, J., & Leuchter, M. (2024). Domain-specific knowledge and domain-general abilities in children's science problem-solving. *British Journal of Educational Psychology*, *94*(2), 346–366. https://doi.org/10.1111/bjep.12649

Schwichow, M., Brandenburger, M., & Wilbers, J. (2022). Analysis of experimental design errors in elementary school: how do students identify, interpret, and justify controlled and confounded experiments? *International Journal of Science Education*, *44*(1), 91–114. https://doi.org/10.1080/09500693.2021.2015544

Sirum, K., & Humburg, J. (2011). The experimental design ability test (EDAT). *Bioscene: Journal of College Biology Teaching*, *37(1)*, 8–16. http://files.eric.ed.gov/fulltext/EJ943887.pdf

Sjøberg, S. (2018). The power and paradoxes of PISA: should inquiry-based science education be sacrificed to climb on the rankings? *Nordic Studies in Science Education*, *14*(2), 186–202. https://journals.uio.no/nordina/article/view/6185/5249

Strat, T. T. S., Henriksen, K. E., & Jegstad, K. M. (2024). Inquiry-based science education in science teacher education: a systematic review. *Studies in Science Education*, *60*(2), 191–249. https://doi.org/10.1080/03057267.2023.2207148

Tairab, H. H. (2015). Assessing students' understanding of control of variables across three grade levels and gender. *International Education Studies*, *9*(1), 44–54. https://doi.org/10.5539/ies.v9n1p44

Teig, N. (2024). Uncovering student strategies for solving scientific inquiry tasks: insights from student process data in PISA. *Research in Science Education*, *54*(2), 205–224. https://doi.org/10.1007/s11165-023-10134-5

van Uum, M. S. J., Verhoeff, R. P., & Peeters, M. (2016). Inquiry-based science education: towards a pedagogical framework for primary school teachers. *International Journal of Science Education*, *38*(3), 450–469. https://doi.org/10.1080/09500693.2016.1147660

van Uum, M. S. J., Verhoeff, R. P., & Peeters, M. (2017). Inquiry-based science education: scaffolding pupils' self-directed learning in open inquiry. *International Journal of Science Education*, *39*(18), 2461–2481. https://doi.org/10.1080/09500693.2017.1388940

Vojíř, K. (2021). What tasks are included in chemistry textbooks for lower- secondary schools: a qualitative view. In M. Rusek, M. Tóthová, & K. Vojíř (Eds.), *Project-based education and other activating strategies in science education XVIII*. https://pages.pedf.cuni.cz/pbe/files/2021/05/ProceedingsPBE2020_final.pdf

Volkmann, M., & Abell, S. (2003). Rethinking laboratories. *Science Teacher*, *70*(6), 38–41. https://www.jstor.org/stable/24156087

World Economic Forum. (2026). *The global risks report 2026*. World Economic Forum. https://reports.weforum.org/docs/WEF_Global_Risks_Report_2026.pdf

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, *92*(5), 941–967. https://doi.org/10.1002/sce.20259

Zetterqvist, A., & Bach, F. (2023). Epistemic knowledge – a vital part of scientific literacy? *International Journal of Science Education*, *45*(6), 484–501. https://doi.org/10.1080/09500693.2023.2166372

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001

Quadrilateral definitions in the Merriam-Webster dictionary: Examining the relationships among quadrilaterals

Samet Okumus[1,*], Tuğrul Kar[1]

[1] Department of Mathematics and Science Education, College of Education, Recep Tayyip Erdoğan University, Cayeli, Rize, 53200, Türkiye; samet.okumus@erdogan.edu.tr

Definitions are of critical importance in both mathematics and mathematics learning. Online dictionaries have become a commonly used resource for students and teachers seeking mathematical definitions. In this context, it is important to examine how mathematical concepts are defined and presented in such publicly accessible platforms. Given its longstanding presence and widespread use, the U.S.-based Merriam-Webster Dictionary was selected for the current study. We examined the definitions of quadrilaterals provided in both the default (Primary) section and the Student Dictionary for Kids. The results indicated that the definitions in the "Kids Definition" section were generally longer and often included superfluous information compared to those in the "Primary" section. Among the special quadrilaterals presented, most were defined in relation to one another, supporting a structure that allows for the identification of subsets within broader quadrilateral types.

Key words: definitions, quadrilaterals, online dictionary, Merriam-Webster Dictionary.

Received 7/2025
Revised 3/2026
Accepted 4/2026

1 Introduction

According to Alcock and Simpson (2017, p. 5), "definitions are central to contemporary formal mathematics because, in order to develop deductive arguments and to communicate clearly, mathematicians need to agree upon precise meanings for mathematical concepts". For instance, definitions specify the distinguishing characteristics of a concept, constitute the foundational components of concept formation, serve as a basis for proof and problem-solving, and facilitate the communication of mathematical ideas by promoting consistency in the meaning of concepts (Zaslavsky & Shir, 2005). For any given mathematical concept, there may be equivalent definitions. The choice among these equivalent definitions may vary depending on personal preference, convenience, or contextual considerations. However, the availability of such options presupposes an awareness of their logical equivalence (Van Dormolen & Zaslavsky, 2003). From a mathematics education perspective, teachers may select definitions based on both pedagogical (e.g., intuitive appeal, alignment with students' needs, clarity, and accessibility for learners) and mathematical (e.g., accuracy) characteristics (Leikin & Winicky-Landman, 2000, 2001).

In the past, individuals may have relied on textbooks, printed dictionaries and encyclopedias to look up mathematical definitions. With the advancement of technology, it has become more common to use online resources, such as artificial intelligence tools and online dictionaries, for the same purpose (Kissane, 2008; Patterson & Young, 2013). Some historical dictionaries now offer online versions to meet new demands, such as increased accessibility. Stakeholders in education (e.g., teachers, students, and parents) may use online dictionaries, for instance, to look up mathematical definitions (Harper et al., 2021; Kissane, 2008; Patterson & Young, 2013). Also, "mathematics textbooks at the elementary school level often follow the standard dictionary conception of definition" (Usiskin & Griffin, 2008, p. 3). So, the influence of dictionaries in mathematics education is widespread. However, writing definitions in mathematics requires expertise in the field. Texts and scripts written by mathematics educators and teachers are not always error-free (Usiskin & Griffin, 2008).

In the present study, we focused on the Merriam-Webster Dictionary – "an Encyclopaedia Britannica company, [that] has been America's leading provider of language information for more than 180 years" (Merriam-Webster, n.d.-a, para. 1). The dictionary includes a general definition section – referred to in this study as the "Primary" definition (Merriam-Webster, n.d.-b) – which appears by default when a word is entered, as well as a "Kids Definition" section (Merriam-Webster, n.d.-c) specifically designed for younger users. We analyzed the definitions of quadrilaterals in both sections, given the longstanding interest in mathematics education research in the precision and clarity of such definitions, as well as in the hierarchical relationships among quadrilaterals (e.g., viewing a square as a specific type of rectangle and defining it as such) (de Villiers, 1994; Jones, 2000; Usiskin & Griffin, 2008).

We concur that, in formal mathematics, it is a general standard that a definition does not contain superfluous information namely, it should be minimal (Usiskin & Griffin, 2008; Zaslavsky & Shir, 2005). Therefore, formal mathematics dictionaries such as the Concise Oxford Dictionary of Mathematics (Clapham & Nicholson, 2009) should adhere to this principle. On the other hand, it is difficult to expect formal mathematical definitions in the Merriam-Webster Dictionary, given its wide intended audience

(Merriam-Webster, 2026). Therefore, one may reasonably encounter colloquial definitions of mathematical terms or characterization-like definitions with superfluous information in dictionaries intended for a broader range of users. Embracing this belief, in the present study, we aimed to address the following research question: In what ways are quadrilaterals defined in the "Primary" and "Kids Definition" sections of the Merriam-Webster Dictionary?

1.1 Rationale for research

According to Zazkis and Leikin (2008), "the definition of a concept, once determined in a curriculum, influences the approach to teaching mathematics, the learning sequence, and the set of theorems and proofs" (pp. 132–133). Much mathematics education research examines definitions of mathematical concepts, particularly quadrilaterals, in textbooks (Abdullah & Shin, 2019; Avcu, 2019; Usiskin & Griffin, 2008). Beyond textbooks, online resources, especially dictionaries, have become increasingly popular for accessing mathematical definitions due to their frequent updates and vast information availability (Kissane, 2008; Patterson & Young, 2013). Given their expanding role in learning, investigating the mathematical and pedagogical characteristics of online dictionaries, addressing limitations, and updating them is vital. However, little research addresses how mathematical concepts are defined in dictionaries.

In the present study, Merriam-Webster is selected as the online dictionary for examining definitions due to its widespread use, high visibility in search engine results, and dual-format offerings. Notably, Merriam-Webster provides both a Primary (general) definition section (Merriam-Webster, n.d.-b) and a "Kids Definition" section tailored for student use (Merriam-Webster, n.d.-c). These sections allow for an analysis of how mathematical ideas are communicated differently to adult and young audiences, which is of pedagogical importance. Since students and mathematics teachers may rely on online searches rather than textbooks for immediate access of definitions, understanding how widely accessible dictionaries present mathematical content is critical. Moreover, as definitions presented in such public resources may shape learners' conceptual understandings, it becomes essential to investigate whether those definitions reflect mathematical accuracy, coherence, and instructional usability. By examining both the mathematical precision and pedagogical clarity of Merriam-Webster's definitions, this study aims to highlight affordances and limitations of such resources in supporting students' understanding of geometric concepts.

Definitions and classifications are interrelated notions in geometric investigations, and together they constitute one of the *big ideas* in school geometry (Sinclair et al., 2012). Classifying and defining geometric objects also provide rich contexts for the development of mathematical reasoning (National Council of Teachers of Mathematics [NCTM], 2000). In this regard, the topic of quadrilaterals provides rich content for both geometric research and the development of mathematical reasoning due to their inclusion of various equivalent statements and alternative classification approaches (de Villiers, 1994; de Villiers et al., 2009; Van Dormolen & Zaslavsky, 2003). Moreover "there is some disagreement in the definitions and, consequently, in the ways in which quadrilaterals are classified and relate to each other" (Usiskin & Griffin, 2008, p. X). Presenting definitions of quadrilaterals to students in classrooms is not a straightforward task for teachers (Avcu, 2023). In this respect, our study, which approaches the definitions and classifications of quadrilaterals in online dictionaries from various perspectives, may contribute to ongoing discussions.

1.2 Theoretical background

As de Villiers et al. (2009, p. 193) emphasized, "a definition that contains conditions (properties) that are both necessary and sufficient is said to be correct". To characterize a correct definition, "it is helpful to recall that logically in the biconditional $p \Leftrightarrow q$, the condition p is viewed as necessary and sufficient for the condition q, meaning that one can conclude that q follows from p, and vice versa" (pp. 193–194). For instance, one may mistakenly define a rhombus as "a convex kite with congruent diagonals that bisect each other" in which a rhombus is depicted as a subset of a kite. However, this definition inadvertently characterizes a special rhombus, namely a square, because having congruent diagonals is not a necessary property of all rhombi. In another example, the definition "a square is a parallelogram with congruent diagonals" includes a necessary property of squares. However, the definition is not sufficient to define a square, as "congruent diagonals" also applies to rectangles with non-congruent adjacent sides. Thus, having congruent diagonals is a necessary but not sufficient condition for defining a square. As is evident, "a definition is incorrect if it contains insufficient or unnecessary properties" (de Villiers, 2009, p. 194).

Simply defined, a quadrilateral is a four-sided polygon (Usiskin & Griffin, 2008). Here, the use of the word 'polygon' connects several essential properties within a single term (Pereira-Mendoza, 1993). Zazkis and Leikin (2008) highlight the use of inaccurate terminology such as a figure and shape when defining special quadrilaterals. For example, according to Zazkis and Leikin (2008, p. 139), defining a square as "a four-sided figure with four equal sides and four equal angles" is an appropriate but non-rigorous example

of a definition. The researchers interpreted that "a 'four-sided figure' or 'shape' implies a polygon", even though such a definition of a square does not make explicit that the sides are composed of coplanar and congruent line segments joined to one another at right angles. Arguably, when sides are referenced in definitions of quadrilaterals, the term evokes polygons. In this context, sides in plane geometry refer to coplanar, conjoined line segments that form a closed figure, thereby implying a polygon – even when non-rigorous terminology is used in definitions of quadrilaterals. Therefore, the use of terms such as shapes or figures in defining quadrilaterals may lead to definitions that are appropriate in context with some non-rigorous examples (Zazkis & Leikin, 2008).

A definition of a special quadrilateral may emphasize different aspects of its properties and establish various relationships among quadrilaterals. For instance, a definition may use minimal properties without including superfluous information. It may also define a special quadrilateral in relation to another, treating it as a subset. In the following subsections, we elaborate on the various aspects of definitions used to characterize relationships among special quadrilaterals.

1.2.1 Partitional and hierarchical definitions among quadrilaterals

The ways in which quadrilaterals have been associated with one another and defined have evolved over the centuries. For instance, in the early development of geometry, Euclid defined a rhombus as "of quadrilateral figures, . . . a rhombus (from Greek: $\rho\mu\beta o\sigma$) that which is equilateral but not right-angled" (Usiskin & Griffin, 2008, p. 19). This definition disjoins squares from rhombi, exemplifying a partitional definition, in which "the concepts involved are considered disjoint from each other" (de Villiers et al., 2009, p. 191). On the other hand, defining a rhombus as "a quadrilateral with congruent sides" establishes a *hierarchical* (inclusive) relationship between a square and a rhombus. Here, the square is accepted as a special case of the rhombus, unlike Euclides' *partitional* (exclusive[1]) definition. As de Villiers (1994, p. 11) emphasized, "by the term hierarchical classification is meant here the classification of a set of concepts in such a manner that the more particular concepts form subsets of the more general concepts". In a similar vein, Usiskin and Griffin (2008, p. 6) pointed out that "an inclusive definition creates a link in a hierarchical chain from the more general to the more specific".

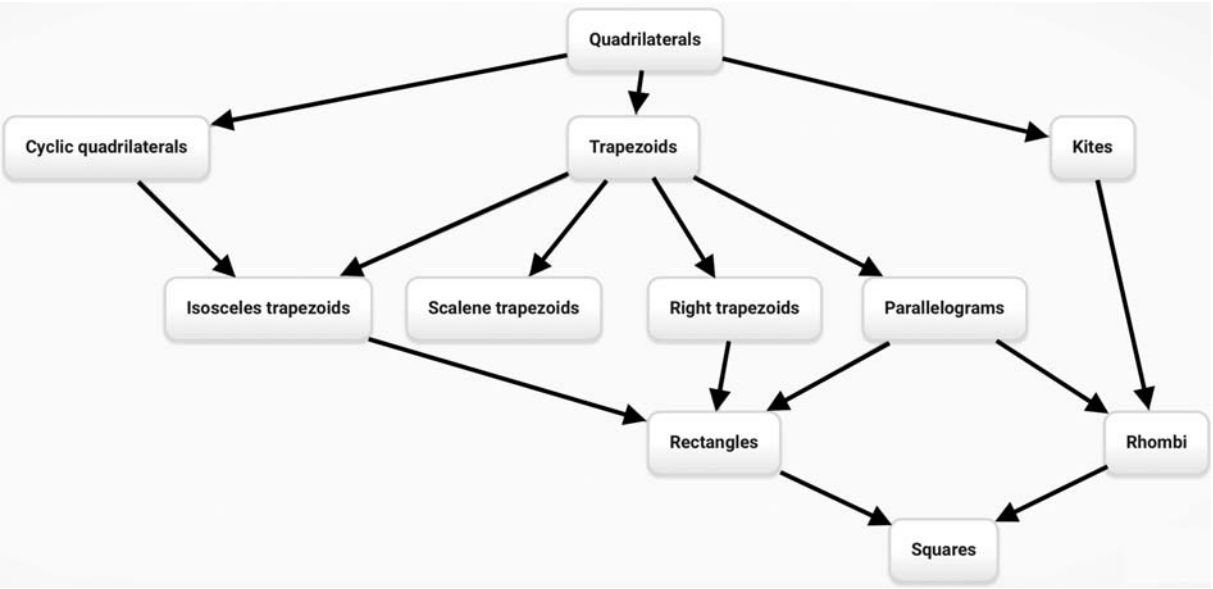

Fig. 1: A hierarchical classification of some quadrilaterals[2]

Partitional definitions can also be made for more general concepts such as quadrilaterals and polygons. For instance, some resources may define a quadrilateral as "the union of four line segments that join four coplanar points, no three of which are collinear, each segment intersecting exactly two others, one at each endpoint" (Usiskin & Griffin, 2008, p. 11). This definition excludes crossed quadrilaterals[3] (de Villiers, 1994). Drawing on Usiskin and Griffin's (2008) study of definition in geometry, Fig. 1 presents a hierarchi-

[1] "An exclusive definition creates a partition of the more general object into a set of more specific objects" (Usiskin & Griffin, 2008, p. 6).

[2] Our classification focuses on convex quadrilaterals, defined as quadrilaterals whose diagonals intersect at a point inside the figure (Graumann, 2005).

[3] "A crossed quadrilateral is a quadrilateral with two of the sides also crossing each other at a point other than the vertices" (de Villiers, 1994, p. 13).

cal classification of the most common special quadrilaterals[4] taught in K–16 mathematics. In Fig. 1, an arrow indicates that the special quadrilateral below is *always* considered a special case of the quadrilateral above. Some mathematics educators also characterize interrelationships between two quadrilaterals using the adverbs *always* and *sometimes* to indicate the hierarchical direction (Duatepe-Paksu & Žilková, 2018; Knapp et al., 2007). For instance, by definition, cyclic quadrilaterals are quadrilaterals whose vertices lie on a circle (Usiskin & Griffin, 2008). Since all isosceles trapezoids, rectangles, and squares can be inscribed in a circle, they are special cases of cyclic quadrilaterals. Let us characterize the hierarchical direction between a cyclic quadrilateral and a rectangle. According to Fig. 1, a cyclic quadrilateral is *sometimes* a rectangle, because cyclic quadrilaterals form a broader class, and only those satisfying the additional condition of having four right angles are rectangles. On the other hand, a rectangle is *always* a cyclic quadrilateral, since the vertices of a rectangle always lie on a circle. The hierarchical classification makes it possible to define a quadrilateral using others that appear higher in the hierarchy (Zazkis & Leikin, 2008).

Some quadrilateral definitions remain debatable, while others are more widely agreed upon. As Usiskin and Griffin (2008, p. 26) emphasized, "there is no disagreement among today's textbook authors regarding which special types of quadrilaterals are always parallelograms. Rectangles, rhombuses, and squares are universally viewed as parallelograms". On the other hand, for a long time, textbooks defined a trapezoid (known as a *trapezium* in British English) as "a quadrilateral with exactly one pair of parallel sides" (Usiskin & Griffin, 2008, p. 27). This definition does not associate a trapezoid with a square, rectangle, rhombus, or parallelogram; therefore, it is a partitional definition.

Some researchers or curriculum writers may still prefer a partitional definition of a trapezoid over a hierarchical one (Casa & Gavin, 2009). However, it is now increasingly common to define a trapezoid hierarchically as "a quadrilateral with at least one pair of opposite parallel sides" (Usiskin & Griffin, 2008, p. 27), making it possible to classify parallelograms, rectangles, rhombi, and squares as special types of trapezoids (Jones, 2000). Using a partitional definition of a trapezoid may not be considered incorrect, but rather idiosyncratic, and "thus the decision one makes in choosing a definition for trapezoid is precisely whether one wishes to include parallelograms in the trapezoid family" (Usiskin & Griffin, 2008, p. 29). Yet, according to de Villiers (1994), hierarchical definitions should be preferred over partitional ones, as they offer both pedagogical and logical advantages.

A hierarchical definition of a special quadrilateral can be expressed concisely, foregrounding only the necessary and sufficient properties. Alternatively, it may include several properties, resulting in a relatively wordy statement. Both types of definitions are common and offer different mathematical and pedagogical affordances. The next section addresses these differences within hierarchical definitions.

1.2.2 Economical and uneconomical hierarchical definitions

In hierarchical definitions that relate quadrilaterals to one another, superfluous properties may be included in addition to those that are necessary and sufficient. As de Villiers et al. (2009) explained, "an economical definition has a minimal set of necessary and sufficient properties; that is, it has no superfluous information. Conversely, an uneconomical definition contains redundant properties" (p. 196). For instance, an uneconomical definition of a rectangle is: "a rectangle is a parallelogram with congruent diagonals and four right angles." The definition is hierarchical, as it also applies to squares, which are considered subsets of rectangles. However, once the right angles are specified, mentioning the congruent diagonals becomes redundant. Therefore, the definition is considered uneconomical.

Zazkis and Leikin (2008) pointed out that the criteria for minimality in definitions remain debatable (see Avcu, 2019; de Villiers et al., 2009; Van Dormolen & Zaslavsky, 2003; Zaslavsky & Shir, 2005). For instance, according to de Villiers et al. (2009, p. 198), defining a rhombus "as a quadrilateral with four congruent sides" is economical "insofar as the defining conditions contain no superfluous information". Yet, for some, referring to the number of sides in the definition is redundant, given that the definition specifies a quadrilateral. In this context, a minimal definition would be: "a rhombus is a quadrilateral with congruent sides." As Usiskin and Griffin (2008, p. 3) underscored, "sometimes authors insert redundant distinguishing characteristics to make it easier for students to deduce properties of the object". On the other hand, according to Zaslavsky and Shir (2005, p. 320) –

> a minimal definition should consist only of information that is strictly necessary for identifying the defined concept. For example, defining a rectangle as a *quadrangle with four right angles* is not a minimal definition, since it is enough to require that there be *three right angles*.

However, this definition requires a deduction to verify that the fourth angle is also a right angle (Jamison, 2000). Thus, while de Villiers et al. (2009) emphasize pedagogical aspects in their approach to

[4]For more types of specials quadrilaterals, see Graumann (2005).

economical definitions, recognizing that different definitions may be preferable for instructional purposes, Zaslavsky and Shir (2005) focus on logical (deductive) aspects, aiming for definitions that are minimal in the strict logical sense. Elsewhere, Van Dormolen and Zaslavsky (2003, p. 96) further elaborated on this distinction as follows:

> On first sight the criterion of minimality seems to be more of an aesthetic or philosophical nature than a logical one. Indeed, describing a rectangle as quadrilateral with four right angles will not result in a contradiction in the system, and may have advantages from a pedagogical perspective. Moreover, often at the time when a concept is defined there may not be sufficient knowledge to determine whether it is minimal. Thus, insisting on this criterion may impede the development of certain concepts or theories.

An economical or uneconomical hierarchical definition may further emphasize different relationships among quadrilaterals both in terms of identifying subsets of the defined concept and in the use of deductions (e.g., logical arguments). The next section elaborates on a fine-grained classification of economical and uneconomical hierarchical definitions.

1.2.3 Subcategorical and deductive relationships within economical and uneconomical hierarchical definitions

An economical or uneconomical definition of a special quadrilateral can be made directly from one of the broadest relevant categories, namely polygons, quadrilaterals or figures/shapes. For instance, a rhombus can be defined as "a quadrilateral with congruent sides and perpendicular diagonals".[5] In this uneconomical definition, the rhombus is defined by means of a general concept of its type, quadrilateral. In another example, the economical definition "a rhombus is a four-sided polygon with congruent sides", uses another general concept 'polygon' and specifies it with its number of sides and properties. Note both definitions are hierarchical, encompassing squares as special cases of rhombi.

Alternatively, what we call a *subcategorical* definition, a special quadrilateral is defined in terms of another special quadrilateral. For instance, a rhombus can be defined by means of kites, parallelograms, or trapezoids (Fig. 1), as in the economical definition: "a rhombus is a trapezoid with congruent sides," or in the uneconomical definition: "a hombus is a parallelogram with perpendicular diagonals that bisect one another." The use of a family relationship within a hierarchy of properties can facilitate "the deductive systemization and derivation of properties of more special concepts" (de Villiers, 1994, p. 15). For example, the subcategorical and economical definition "a square is a rhombus with a right angle" requires establishing that one right angle in a rhombus necessitates that the other three angles are also right angles (Jamison, 2000). When students encounter this definition, they may question the measures of the rest of the angles in the rhombus, since the definition refers to *a right angle*. According to Jamison (2000), such definitions should be dispreferred. Regarding the subcategorical-economical definition "a rectangle is a *parallelogram* whose diagonals have equal lengths" (Jamison, 2000, p. 49), Jamison claimed that the statement is in the form of a theorem to be proved. He posited:

> This statement is true and concise, but the defining property is not BASIC. This would work better as a theorem to be proved than as a definition. In mathematics, assertions of this kind are regarded as *characterizations* rather than as definitions.

Jamison (2000, p. 48) contended that "a definition is a *concise* statement of the *basic* properties of an object or concept which *unambiguously identify* that object or concept... It should involve *basic* properties, ideally those that are simply stated and have immediate intuitive appeal". In this context, Jamison does not appreciate theorem-like definitions "that require extensive derivation or are hard to work with" such as using diagonal properties (e.g., diagonals that bisect each other, and perpendicular diagonals) to deduce congruent sides or angles in a quadrilateral. Then, making a subcategorical definition may be challenging for students, since it necessitates associating a special quadrilateral with another one, which increases complexity of definitions compared to defining a special quadrilateral by means of a polygon or a quadrilateral (Jamison, 2000).

On the other hand, de Villiers et al. (2009, p. 199) referred to convenient definitions, stating: "a good, or convenient, definition is one that also allows us to deduce the other properties of the concept easily; that is, it should be *deductive-economical*". Deductive properties necessitate the use of theorems or other properties (e.g., congruence), requiring proofs or arguments in quadrilateral definitions (de Villiers et al., 2009; Van Dormolen & Zaslavsky, 2003). For example, de Villiers et al. (2009, p. 199) provided the following definition of a rhombus as "a quadrilateral with two axes of symmetry through the opposite

[5] Definitions in quotation marks without citations are those provided by the authors of this study.

angles". This economical definition of a rhombus draws on symmetry, which must relate to congruency in order to deduce its properties. Consistent with this view, Usiskin and Griffin (2008, p. 57) stated the theorem "every rhombus is symmetric to the lines containing its diagonals", thereby treating symmetry as a consequence to be proved, rather than as an assumed defining property.

In a similar vein, other types of definitions (e.g., subcategorical, uneconomical) may also possess deductive qualities. For instance, the definition "a rectangle is an isosceles trapezoid with congruent diagonals that bisect each other" includes superfluous information – specifically, the term "congruent" is redundant, as an isosceles trapezoid always has congruent diagonals. Therefore, this is considered an uneconomical definition. Additionally, it is subcategorical, as it defines a rectangle by associating it with an isosceles trapezoid. This definition also requires the use of the diagonal properties of the isosceles trapezoid in the sense of diagonal bisection – a property not shared by all isosceles trapezoids. Therefore, it involves deduction and is thus characterized as *deductive–subcategorical–uneconomical*.

Overall, Fig. 2 presents a chart that depicts the types of hierarchical definitions, accompanied by sample definitions within the classification of quadrilaterals. A hierarchical definition for quadrilaterals can be formulated either economically or uneconomically. An economical definition from a *pedagogical* perspective (de Villiers et al., 2009) uses the necessary and sufficient properties, though it is not necessarily strictly minimal from a logical perspective, as emphasized by Van Dormolen and Zaslavsky (2003). In contrast, an uneconomical definition includes superfluous information, even though it still contains the necessary and sufficient properties of the defined concept (de Villiers et al., 2009). If an economical or uneconomical definition uses one special quadrilateral to define another, it is further characterized as a subcategorical definition. If the definition involves deducing properties (e.g., through proofs or logical arguments), it is further categorized as a deductive definition. Finally, if an economical or uneconomical definition incorporates both aspects, it is considered both subcategorical and deductive.

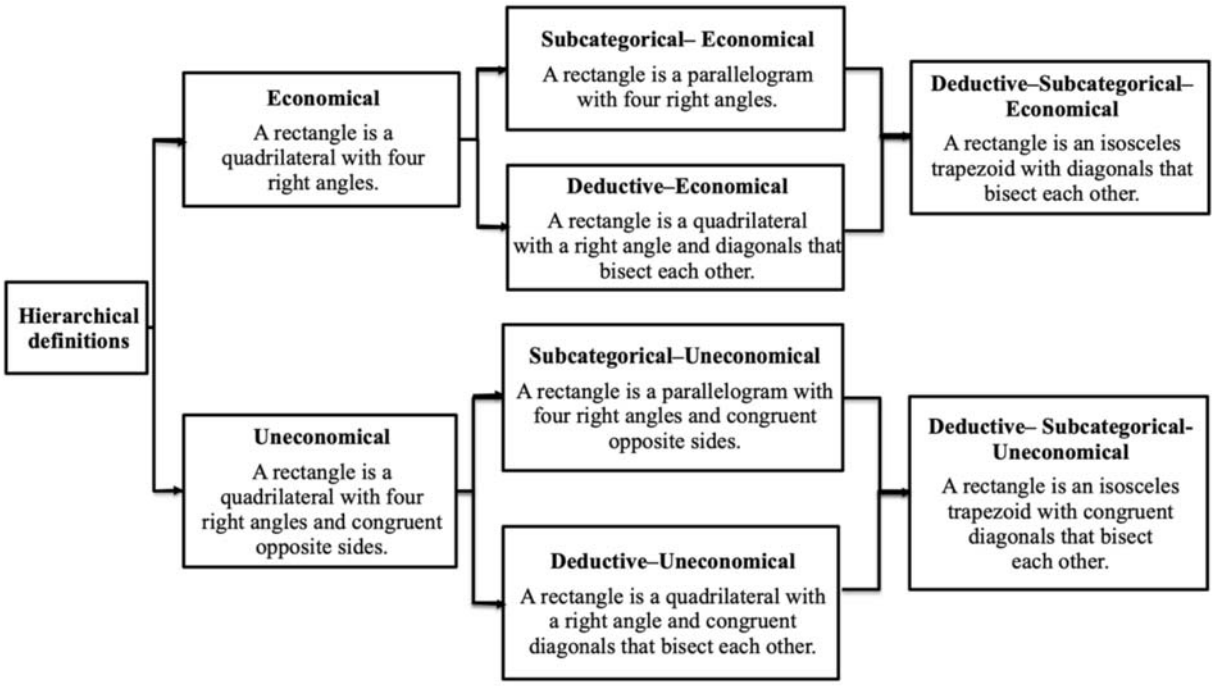

Fig. 2: Types of hierarchical definitions and sample definitions

2 Methods

We first searched for popular online dictionaries and excluded commercial dictionaries (e.g., Oxford English Dictionary) that do not provide free access to users. We selected the U.S.-based Merriam-Webster Dictionary due to its online accessibility and long-standing historical presence (Merriam-Webster, n.d.-a). From a hierarchical classification perspective (Fig. 1), we examined the most common special quadrilaterals taught in K–16 mathematics. In this context, we searched for the following geometric shapes: cyclic quadrilateral, isosceles trapezoid, kite, parallelogram, quadrilateral, rectangle, rhombus, right trapezoid, scalene trapezoid, square, trapezium, and trapezoid. However, the *mathematical* definitions found in the dictionary was limited to rhombus, parallelogram, quadrilateral, rectangle, rhombus, square, trapezium, and trapezoid.

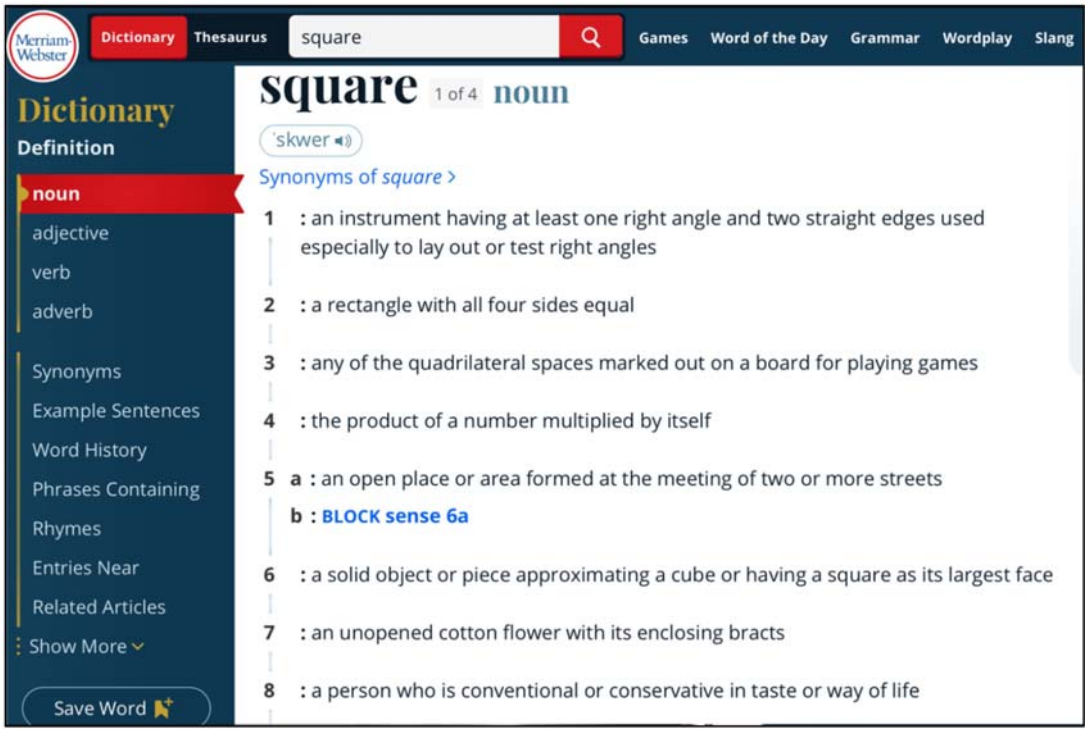

Fig. 3: Multiple noun definitions of square in the Primary Definition section

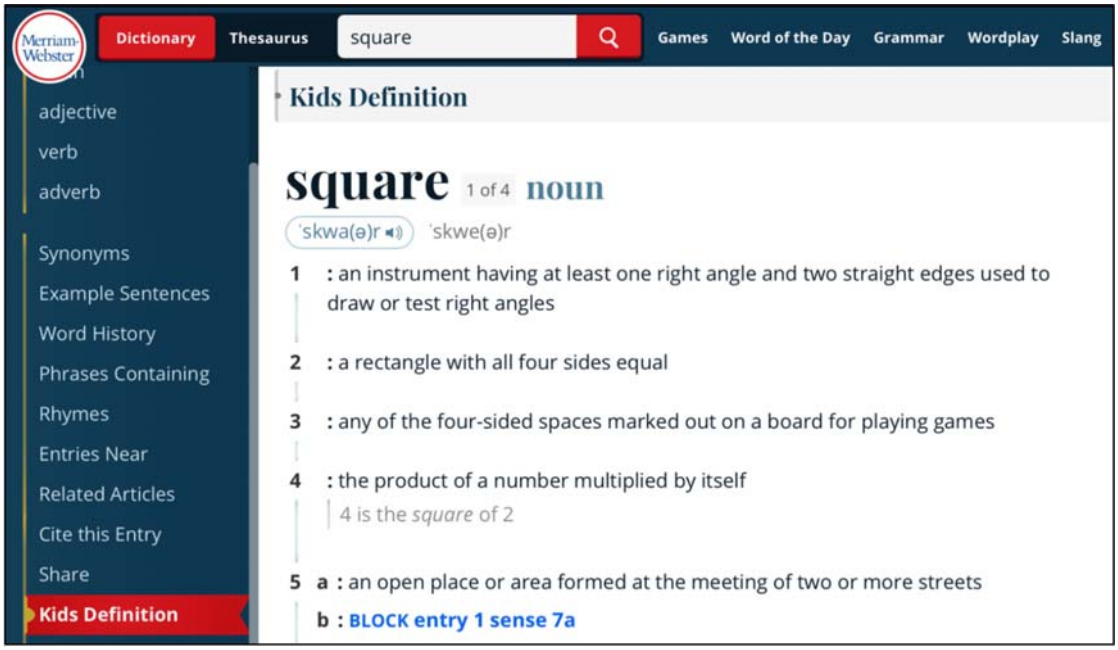

Fig. 4: Multiple noun definitions of square in the "Kids Definition" section

When a word is looked up in the Merriam-Webster Dictionary, the "Primary" Definition section (Merriam-Webster, n.d.-b) provides its noun, adjective, verb, and adverb meanings, if applicable (Fig. 3). For our study, we examined the mathematical definitions listed under the noun form. Among the noun definitions, we identified mathematics-related definitions of the concept when multiple meanings existed (e.g., colloquial uses). For instance, Fig. 3 presents several definitions of a square. Only the second definition pertains to the quadrilateral. The sixth definition, although related to mathematics, concerns solid geometry and was therefore excluded from our study.

The dictionary also includes a "Kids Definition" section, described as: "Search an online dictionary written specifically for young students. Kid-friendly meanings from the reference experts at Merriam-Webster help students build and master vocabulary" (Merriam-Webster, n.d.-c, para. 1). We also searched for quadrilaterals (e.g., Fig. 4) in the "Kids Definition" section and applied a similar selection process to identify relevant definitions for our study, as was done in the "Primary" definition section.

2.1 Coding and data analysis

We analyzed the mathematical definitions of rhombus, parallelogram, quadrilateral, rectangle, rhombus, square, trapezium, and trapezoid in both sections of the Merriam-Webster Dictionary. Since the online definitions may change over time, we provided screenshots of the analyzed definitions in our data presentation. Each definition was evaluated to determine whether it accurately provided the necessary and sufficient conditions for the geometric shape. Definitions that failed to meet these conditions were categorized as incorrect. If a definition disjointed a special quadrilateral from others, it was coded as partitional. Finally, if a definition provided the necessary and sufficient properties, leading to hierarchical relationships among quadrilaterals, it was classified as hierarchical.

Hierarchical definitions were further categorized based on whether they contained superfluous information; in particular, we identified whether they were formulated economically or uneconomically. We adopted the principle of minimality in definitions, as advocated by de Villiers et al. (2009), emphasizing pedagogical advantages rather than aesthetic or philosophical considerations, as discussed by Van Dormolen and Zaslavsky (2003). Our decision to adopt a pedagogical perspective was particularly influenced by our analysis of the "Kids Definition" section.

Finally, we examined whether an economical or uneconomical hierarchical definition incorporated deductive properties and whether one type of special quadrilateral was used to define another. For instance, a rhombus can be defined by means of a special quadrilateral (trapezoid, parallelogram, or kite) or a more general concept (e.g., quadrilateral, polygon, or geometric shape). If it is defined by means of a special quadrilateral, we coded it as subcategorical, as in the following economical definition: "a rhombus is a parallelogram with adjacent congruent sides." Therefore, the definition is subcategorical-economical.

In another example, the economical definition "a rhombus is a quadrilateral with at least two lines of symmetry" incorporates a deduction-based argument, as it entails applying a reflection transformation to quadrilaterals and deducing that any quadrilateral with two or four lines of symmetry meets the criteria, given the requirement of "at least two lines of symmetry." Therefore, it is categorized as deductive-economical. If both deductive and subcategory-based elements are present in an economical definition, as in the definition "a rhombus is a kite with at least two lines of symmetry," it is categorized as deductive-subcategorical-economical. A similar coding scheme was applied to uneconomical definitions.

Table 1 presents the codebook used in our data analysis. We categorized each quadrilateral according to the coding scheme, and constructed data matrices for both sections of the dictionary. In total, we

Table 1: Codebook of data analysis

Definition type	Description	Sample definition
Incorrect	The definition does not have the necessary and/or sufficient properties of the concept.	A rhombus is a parallelogram with all right angles.
Partitional	The definition disjoints a special quadrilateral from other types, not leading to hierarchical relationships among quadrilaterals.	A rhombus is a quadrilateral with equal side lengths and no right angles.
Hierarchical		
Uneconomical	The definition with the necessary and sufficient properties of the concept provides superfluous information.	A rhombus is a quadrilateral with congruent sides and perpendicular diagonals.
Economical	The definition with the necessary and sufficient properties of the concept provides no superfluous information.	A rhombus is a quadrilateral with congruent sides.
Deductive-Uneconomical	An uneconomical definition that also allows for deducing properties of the concept by using theorems or other properties, requiring proofs or logical arguments in quadrilateral definitions.	A rhombus is a quadrilateral with perpendicular diagonals and at least two lines of symmetry.
Deductive-Economical	An economical definition that also allows for deducing properties of the concept by using theorems or other properties, requiring proofs or logical arguments in quadrilateral definitions.	A rhombus is a quadrilateral with at least two lines of symmetry.
Subcategorical-Uneconomical	The definition establishes a classification in which a more specific concept is associated with a more general one with superfluous information.	A rhombus is a parallelogram with adjacent congruent sides and perpendicular diagonals.
Subcategorical-Economical	The definition establishes a classification in which a more specific concept is associated with a more general one with no superfluous information.	A rhombus is a parallelogram with adjacent congruent sides.
Deductive-Subcategorical-Uneconomical	A subcategorical-uneconomical definition that also allows for deducing properties of the concept by using theorems or other properties, requiring proofs or logical arguments in quadrilateral definitions.	A rhombus is a kite with adjacent congruent sides and at least two lines of symmetry.
Deductive-Subcategorical-Economical	A subcategorical-economical definition that also allows for deducing properties of the concept by using theorems or other properties, requiring proofs or logical arguments in quadrilateral definitions.	A rhombus is a kite with at least two lines of symmetry.

coded seven definitions in the "Primary" section and six definitions in the "Kids Definition" section. First, the authors coded the definitions in the "Kids Definition" section individually. Then, the authors met and compared their coding, noting any differences. In this process, the coding scheme was revisited to ensure consistent interpretation and application of the categories through ongoing discussion. For instance, a rhombus is defined as "a parallelogram with all four sides of equal length and usually with no right angles" in the "Kids Definition" section. One of the authors thought this definition suggested a relatively partitional definition since the adverb 'usually' was used, which might imply not associating a rhombus with a square. The other author pointed out that the definition did not prevent classifying a square as a subset of a rhombus. The authors reached a consensus that the use of the adverb 'usually' did not exclude a square from being classified as a rhombus. Therefore, the definition was not categorized as partitional. A similar coding negotiation was used in the "Primary" section of the dictionary. Finally, we compared the definitions across the "Primary" and "Kids Definition" sections.

3 Results

Across the dictionary sections, definitions of cyclic quadrilaterals, kites, isosceles trapezoids, and scalene trapezoids are not presented in either section. Table 2 summarizes the types of definitions found in the "Primary" and "Kids Definition" sections of the dictionary. A partitional definition of 'trapezoid' is presented in both sections. In the "Primary" section, 'trapezoid' is defined as "a quadrilateral having only two sides parallel," while the Kids Definition section states, "a polygon that has four sides and exactly two that are parallel" (Fig. 5). Notably, the latter definition is longer and defines a trapezoid in terms of a polygon. In both cases, 'trapezoid' is defined using partitional classifications, excluding the parallelogram family from being considered subsets of trapezoids. Moreover, the definition of 'trapezium' in the "Primary" section as "a quadrilateral with no sides parallel" is incorrect.

Table 2: Definition types across the dictionary sections

Types of definitions	Primary	Kids Definition
Incorrect	Trapezium	–
Partitional	Trapezoid	Trapezoid
Hierarchical		
Uneconomical	Parallelogram	Parallelogram, Rectangle, Quadrilateral
Economical	Quadrilateral	–
Subcategorical-Uneconomical	Rhombus	Rhombus
Subcategorical-Economical	Square, Rectangle	Square

Fig. 5: The definitions of trapezoid/trapezium across the sections of the dictionary

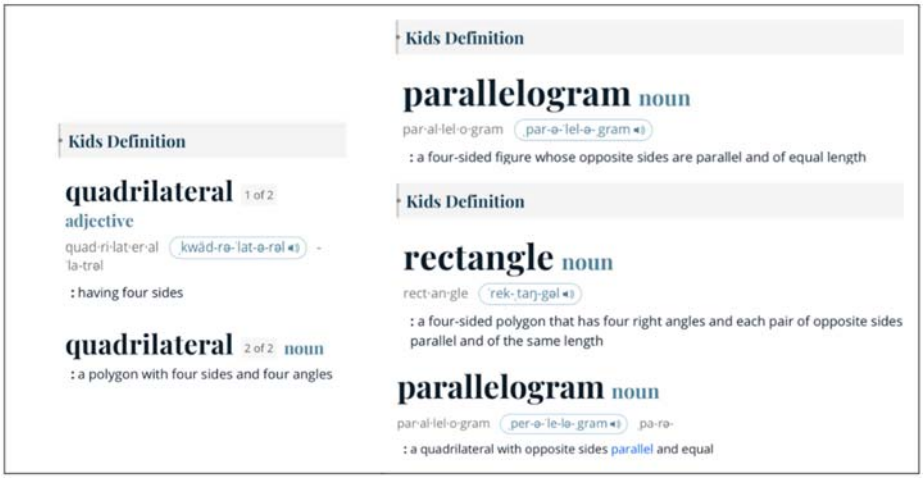

Fig. 6: The uneconomical definitions

The definition of 'parallelogram' in the "Primary" section, and those of parallelogram, rectangle, and quadrilateral in the Kids Definition section, are presented uneconomically with superfluous information (Fig. 6). The "Primary" definition of 'parallelogram' reads: "a quadrilateral with opposite sides parallel

and equal." Here, once the parallelism of opposite sides is stated, referring to their equality in length becomes redundant. Additionally, there is a minor language issue in this definition, as the term congruent would be more appropriate than equal when referring to geometric figures.

In the Kids Definition section, the uneconomical definitions of 'quadrilateral' and 'rectangle' use the term polygon in their descriptions (Fig. 6). For example, in the definition of 'quadrilateral,' after stating the number of sides, mentioning the number of angles is superfluous. The uneconomical definition of 'parallelogram' in the Kids Definition section reads: "a four-sided figure whose opposite sides are parallel and of equal length." Although appropriate in context, this definition is non-rigorous, as the term figure implies a polygon, but polygonal properties (e.g., closed and coplanar sides) are not explicitly stated.

Different from the uneconomical definition provided for 'quadrilateral' in the Kids Definition section, 'quadrilateral' is defined economically, with no superfluous information, in the "Primary" section (Fig. 7). The definition is based on a polygon with four sides.

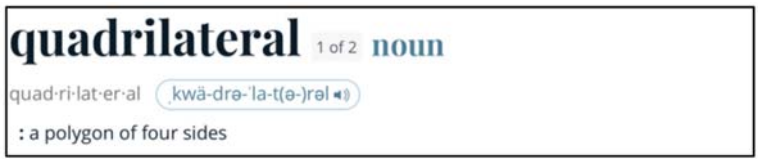

Fig. 7: Quadrilateral definition in the Primary section

Fig. 8: Subcategorical-uneconomical definitions

'Rhombus' in both sections is defined using subcategorical-uneconomical definitions (Fig. 8). The definition of 'rhombus' is given by means of a parallelogram, and it reads: "a parallelogram with four equal sides and sometimes one with no right angles." Here, after referring to equal side lengths, the mention of a parallelogram with non-right angles is redundant. From a hierarchical point of view, by adding "sometimes one with no right angles" to the definition, the statement implies a parallelogram with congruent sides is *sometimes* a non-right-angled rhombus, despite the fact that a parallelogram with congruent sides may have all right angles. In addition, the definition presents a minor language issue, as it uses "equal sides" instead of referring more precisely to "equal side lengths" or "congruent sides." In the Kids Definition section, 'rhombus' is also defined by means of a parallelogram: "a parallelogram with all four sides of equal length and usually with no right angles." Here, the definition contains superfluous information in its reference to parallelograms "usually with no right angles." Notably, the adverb 'usually' is used, instead of 'sometimes.' This word of choice suggests a preference for associating a parallelogram with a non-right-angled rhombus, while treating the case of a square as less preferred or implicitly dispreferred, despite the fact that a square is also a rhombus within the same hierarchical structure.

Finally, the definitions of 'rectangle' and 'square' in the "Primary" Dictionary, and 'square' in the "Kids Definition" section, are subcategorical-economical (Fig. 9). The definition of 'rectangle' is given by means of a parallelogram and does not contain superfluous information from a pedagogical viewpoint. It is important to note that the definition includes the clarification "*especially*: one with adjacent sides of unequal length," which is not part of the core definition. This clarification does not establish a partitional classification between squares and rectangles. Yet it may subtly reinforce the perception of squares as conceptually different from rectangles, despite their inclusion within the definition as a type of rectangle. The definition of 'square' in both sections is stated as "a rectangle with all four sides equal." This definition uses a special quadrilateral (rectangle) with minimal properties; therefore, it is subcategorical-

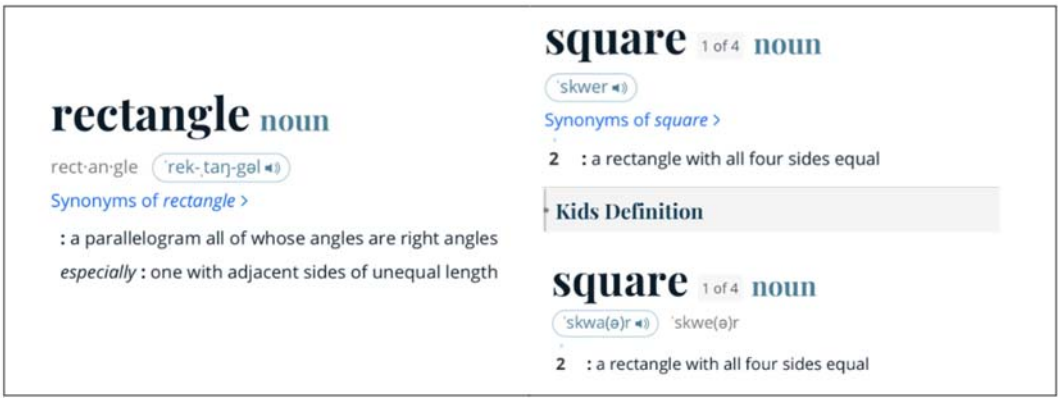

Fig. 9: Subcategorical-economical definitions

economical. A similar language issue appears in the square definition regarding the use of the term equal. More precise wordings could include: four congruent sides, four equal side lengths, or four sides of equal length.

4 Discussion

Online dictionaries have increasingly influenced mathematics education due to their easy access to vast information (Kissane, 2008; Patterson & Young, 2013). While they offer clear benefits for students and teachers, their limitations and areas for improvement warrant attention. Addressing these can raise awareness among practitioners and guide dictionary providers in enhancing their services. Considering the complexities and disagreements surrounding quadrilateral definitions and their instructional challenges (Avcu, 2023; de Villiers, 1994; Usiskin & Griffin, 2008), this study examines in what ways quadrilaterals are defined in the Merriam-Webster Dictionary's "Primary" and "Kids Definition" sections.

We identified seven definitions related to quadrilaterals in the "Primary" section and six in the "Kids Definition" section of the dictionary. Notably, the definition of 'trapezium' is inaccurate, and the definition of 'trapezoid' is framed using a partitional approach. Beyond analyzing the formal aspects of the "Primary" and "Kids Definition" sections in the Merriam-Webster Dictionary (i.e., the types of quadrilaterals included), our study also provides significant insights into the mathematical content of these definitions. A key criterion for evaluating definitions is their accuracy, which involves whether they adequately capture the necessary and sufficient properties of the geometric object (de Villiers et al., 2009).

Hierarchical definitions describe a quadrilateral in terms of another, thereby facilitating an understanding of the relationships among different quadrilaterals (Zazkis & Leikin, 2008). In this regard, they should be preferred over partitional definitions due to their pedagogical and logical advantages (de Villiers, 1994). Our results indicate that both sections of the Merriam-Webster Dictionary predominantly employ hierarchical definitions, which aligns well with the types of definitions taught within the K–16. However, a hierarchical definition for the trapezoid is absent, and thus, the dictionary does not represent the parallelogram family as a subset of trapezoids. Although this omission does not constitute an outright error, revising such definitions to better align with the curricula and textbooks could be considered to enhance consistency and educational relevance.

Two definitions in the "Primary" definitions section (parallelogram and rhombus) and four in the "Kids Definition" section (parallelogram, rectangle, quadrilateral, and square) are identified as uneconomical. The greater level of detail in the "Kids Definition" section, intended to support students' understanding of quadrilaterals, may have contributed to the higher incidence of uneconomical definitions. As addressed earlier, uneconomical definitions in mathematics are often presented in the form of characterizations with superfluous information to help especially young students understand the terms (Jamison, 2000; Usiskin & Griffin, 2008). The publisher (Merriam-Webster, 2026) might have had a similar motive in using characterization-like statements for their readers. Also, in both the "Primary" and "Kids Definition" sections, two approaches to hierarchical definitions are evident. Some specific quadrilaterals are described using broad terms like polygons and shapes (e.g., parallelogram in the Kids section), which risks omitting essential properties and may be pedagogically discouraged (Pereira-Mendoza, 1993). Others are defined by means of special quadrilaterals (i.e., subcategorical definitions), supporting deductive systematization and the derivation of properties related to more specific concepts (de Villiers, 1994). Notably, some definitions have issues with mathematical language, particularly regarding the use of "equal"

to describe the relationships between side lengths, where more precise terminology such as "equal side lengths" or "congruent sides" would have been appropriate. It should be noted that providing definitions with imprecise language may impede students' deductive reasoning in the process of creating and critiquing geometric ideas relationships (NCTM, 2000).

Another important result is the absence of deductive definitions in both sections of the dictionary. These definitions, which involve deriving properties through proofs or other geometric attributes (de Villiers et al., 2009; Van Dormolen & Zaslavsky, 2003), can also be embedded within the subcategorial definition types, though they may result in more complex description. Considering that online dictionaries cater to a broad audience with varying mathematical backgrounds, a deductive definition may not have been provided in a lay language, as such definitions presuppose familiarity with additional mathematical properties and relationships that may exceed the background knowledge of many users.

5 Final remarks

Our results revealed that some commonly used types of quadrilaterals in K–16 education (see Fig. 1) were not present in the online dictionary, and even among the defined types, discrepancies were observed between the "Primary" and "Kids Definition" sections. Definitions of cyclic quadrilaterals, kites, isosceles trapezoids, and scalene trapezoids are absent from both sections of the dictionary. While all these quadrilateral types are incorporated within K–16 mathematics, some are introduced in upper grade levels, such as high school (National Governors Association Center for Best Practices, Council of Chief State School Officers [NGA & CCSSO], 2010; NCTM, 2000). The Merriam-Webster Dictionary does not specify the age group targeted by its "Kids Definition" section. Assuming that this section targets young students, the omission of definitions for terms such as cyclic quadrilaterals and kites can be justified on developmental or curricular grounds. However, considering that the "Primary" definitions section is intended for all users and that over 60% of the dictionary's users are aged between 18 and 34 (Merriam-Webster, 2026), it is important that this section comprehensively encompasses all the common hierarchical definitions.

Enriching the dictionary with definitions of various types of quadrilaterals not only meets the needs of diverse users but also reinforces its role as a reliable and accessible resource for both students and professionals. While the Merriam-Webster Dictionary is not a specialized mathematics dictionary, it plays an influential role in shaping public understanding of mathematical terminology, especially due to its wide accessibility and prominence in online search results. According to the publisher –

> Each day most Merriam-Webster editors devote an hour or two to reading a cross section of published material, including books, newspapers, magazines, and electronic publications; in our office this activity is called "reading and marking." ... Any word of interest is marked, along with surrounding context that offers insight into its form and use. (Merriam-Webster, n.d.-d, para. 4)

Given this editorial process, definitions of additional special quadrilaterals may be incorporated into the dictionary in the future. Such updates would inevitably raise important questions about how these geometric figures should be defined. One of the considerations in this regard is the recognition of hierarchical relationships among quadrilaterals. Another observation from our analysis concerns the relative brevity of the definitions in the "Primary" section of the Merriam-Webster Dictionary, which are often shorter than those in the "Kids Definition" section.

While conciseness can be valuable, our findings suggest that using fewer words does not necessarily enhance clarity or precision. In some cases, essential properties of geometric figures are omitted, leading to conceptual ambiguities. Consider, for instance, the following definition: "An isosceles trapezoid is a trapezoid in which at least one pair of opposite sides are congruent" (Usiskin & Griffin, 2008, p. 42). According to this definition, a parallelogram with non-right angles could be interpreted as a special case of an isosceles trapezoid. However, not all parallelograms are symmetric. Therefore, they do not satisfy the formal properties typically associated with isosceles trapezoids. In such definitions, the use of symmetry in the definition is essential to avoid misclassification and ensure mathematical accuracy. An accurate hierarchical definition would state: "An isosceles trapezoid is a trapezoid that is symmetric about at least one line passing through the midpoints of its opposite sides." Such definitions underscore the importance of ensuring that dictionary definitions, especially those accessible to a broad audience, align with both mathematical rigor and pedagogical clarity.

In this study, we focused on the ways in which definitions of quadrilaterals are presented in the Merriam-Webster Dictionary. However, our study does not investigate how definitions in the Merriam-Webster Dictionary are used within teaching and learning context. Consequently, the present study falls short in terms of the practical use of the dictionary, as it does not reflect on how stakeholders

in education interpret dictionary definitions in classroom instruction, tutoring practices, or informal learning settings. In this context, further research is needed to explore how, for instance, students make sense of mathematical definitions encountered in online dictionaries, how these definitions are examined and negotiated within classroom discourse, and how they are aligned with the goals of instruction. Such investigations have the potential not only to inform the refinement of mathematical definitions in digital resources, but also to enhance our understanding of the role these resources may play in shaping teaching and learning processes. As online dictionaries continue to evolve, collaboration with mathematicians and mathematics educators may become increasingly essential to prevent the perpetuation of misconceptions and to support learners at all levels.

References

Abdullah, H. A., & Shin, B. (2019). A comparative study of quadrilaterals topic content in mathematics textbooks between Malaysia and South Korea. *Journal on Mathematics Education*, *10*(3), 315–340.

Alcock, L., & Simpson, A. (2017). Interactions between defining, explaining and classifying: The case of increasing and decreasing sequences. *Educational Studies in Mathematics*, *94*(1), 5–19. https://doi.org/10.1007/s10649-016-9709-4

Avcu, R. (2019). A comparison of mathematical features of Turkish and American textbook definitions regarding special quadrilaterals. *International Journal of Mathematical Education in Science and Technology*, *50*(4), 577–602. https://doi.org/10.1080/0020739X.2018.1529338

Avcu, R. (2023). Pre-service middle school mathematics teachers' personal concept definitions of special quadrilaterals. *Mathematics Education Research Journal*, *35*(4), 743–788. https://doi.org/10.1007/s13394-022-00412-2

Casa, T. M., & Gavin, M. K. (2009). Advancing elementary school students' understanding of quadrilaterals. In T. Craine. & R. Rubenstein (Eds.), *Understanding Geometry for a Changing World. 71st Yearbook* (pp. 205–219). NCTM.

Clapham, C., & Nicholson, J. (2009). *The concise Oxford dictionary of mathematics.* Oxford University Press.

de Villiers, M. (1994). The role and function of a hierarchical classification of the quadrilaterals. *For the Learning of Mathematics*, *14*(1), 11–18.

de Villiers, M., Govender, R., & Patterson, N. (2009). Defining in Geometry. In T. Craine & R. Rubenstein (Eds.), *Understanding Geometry for a Changing World. 71st Yearbook* (pp. 189–204). NCTM.

Duatepe-Paksu, A., & Žilková, K. (2018). The content knowledge of Turkish and Slovak pre-service elementary teachers: the case of the square. In P. Hájek & O. Vít (Eds.), *CBU International Conference Proceedings* (Vol. 6, pp. 556–561). CBU Research Institute. https://doi.org/10.12955/cbup.v6.1213

Graumann, G. (2005). Investigating and ordering Quadrilaterals and their analogies in space—problem fields with various aspects. *ZDM*, *37*(3), 190–198. https://doi.org/10.1007/s11858-005-0008-2

Harper, F. K., Rosenberg, J. M., Comperry, S., Howell, K., & Womble, S. (2021). #Mathathome during the COVID-19 pandemic: Exploring and reimagining resources and social supports for parents. *Education Sciences*, *11*(2), 1–24. https://doi.org/10.3390/educsci11020060

Jamison, R. E. (2000). Learning the language of mathematics. *Language and Learning across the Disciplines*, *4*(1), 45–54. https://doi.org/10.37514/LLD-J.2000.4.1.06

Jones, K. (2000). Providing a foundation for deductive reasoning: Students' interpretation when using dynamic geometry software and their evolving mathematical explanations. *Educational Studies in Mathematics*, *44*(1–3), 55–85. https://doi.org/10.1023/A:1012789201736

Kissane, B. (2008). Learning mathematics on the internet. In T. de Alwis, M. Majewski & W.-C. Yang (Eds.), *the 13th Asian Technology Conference in Mathematics.* https://atcm.mathandtech.org/ep2008/papers_full/2412008_15110.pdf

Knapp, A. K., Barret, J. E., & Kaufmann, M. L. (2007). Prompting teacher knowledge development using dynamic geometry software. In T. Lamberg & L. Wiest (Eds.), *Proceedings of the 29th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1098–1105). University of Nevada, Reno.

Leikin, R., & Winicky-Landman, G. (2000). On equivalent and nonequivalent definitions: Part 2. *For the Learning of Mathematics*, *20*(2), 24–29.

Leikin, R., & Winicky-Landman, G. (2001). Defining as a vehicle for professional development of secondary school mathematics teachers. *Mathematics Teacher Education and Development*, *3*, 62–73.

Merriam-Webster. (n.d.-a). *About us – Merriam-Webster Dictionary.* Retrieved June 24, 2025, from https://www.merriam-webster.com/about-us

Merriam-Webster. (n.d.-b). *Dictionary – Merriam-Webster Dictionary.* Retrieved June 24, 2025, from https://www.merriam-webster.com/dictionary/

Merriam-Webster. (n.d.-c). *Student dictionary for kids – Merriam-Webster Dictionary.* Retrieved June 24, 2025, from https://www.merriam-webster.com/kids

Merriam-Webster. (n.d.-d). *How does a word get into a Merriam-Webster dictionary?* Retrieved June 24, 2025, from https://www.merriam-webster.com/help/faq-words-into-dictionary

Merriam-Webster. (2026). *Advertising – Merriam-Webster Dictionary.* Retrieved January 19, 2026, from https://www.merriam-webster.com/assets/mw/static/pdf/advertising/MWEB_Media_Kit_2017.pdf

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* NCTM.

National Governors Association Center for Best Practices, Council of Chief State School Officers [NGA & CCSSO]. (2010). *Common core state standards for mathematics.* http://www.corestandards.org

Patterson, L. G., & Young, A. F. (2013). The power of math dictionaries in the classroom. *SRATE Journal*, *22*(2), 22–28.

Pereira-Mendoza, L. (1993). What is a quadrilateral? *The Mathematics Teacher*, *86*(9), 774–776.

Sinclair, N., Pimm, D., Skelin, M., & Zbiek, R. M. (2012). *Developing essential understanding of geometry: Grades 6–8.* NCTM.

Usiskin, Z., & Griffin, J. (2008). *The classification of quadrilaterals: A study of definition. A Volume in Research in Mathematics Education.* Information Age Publishing.

Van Dormolen, J., & Zaslavsky, O. (2003). The many facets of a definition: The case of periodicity. *The Journal of Mathematical Behavior*, *22*(1), 91–106. https://doi.org/10.1016/S0732-3123(03)00006-3

Zaslavsky, O., & Shir, K. (2005). Students' conceptions of a mathematical definition. *Journal for Research in Mathematics Education*, *36*(4), 317–346.

Zazkis, R., & Leikin, R. (2008). Exemplifying definitions: a case of a square. *Educational Studies in Mathematics*, *69*(2), 131–148. https://doi.org/10.1007/s10649-008-9131-7

Using a large language model to analyse problem solving and simulate student solutions in lower secondary mathematics

Jiří Přibyl[1,*], Michaela Tichá[1]

[1] Faculty of Science, Jan Evangelista Purkyně University, Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic; jiri.pribyl@ujep.cz

This study examines how a large language model (LLM) can support lower secondary mathematics education, not by measuring performance, but by analysing how complete and didactically useful its solution processes are. The focus is on whether the model explicitly verifies results, which is an important but often neglected phase of mathematical problem solving. Two exploratory experiments were conducted with Gemini 2.5 Pro using 24 lower secondary mathematics tasks presented in Czech. In Experiment 1, the model solved all tasks in a baseline condition and then again using a structured prompt inspired by Pólya's four phases of problem solving. In Experiment 2, the model was asked to simulate solution attempts of three contrasting student profiles. The results show that verification is a fragile step. In the baseline condition, the model often produced correct answers but did not check them. When guided by the structured prompt, verification appeared in every task. In student simulation, the model produced plausible mistakes and omissions in routine tasks, but often became unrealistically advanced in non-routine and construction problems, reducing profile fidelity. Overall, LLM outputs can support a priori didactic analysis, but they require careful interpretation, especially when used to simulate student thinking in demanding tasks.

Key words: mathematical problem solving, result verification, a priori didactic analysis, large language models, lower secondary education.

Received 2/2026
Revised 4/2026
Accepted 4/2026

1 Introduction

1.1 Mathematical problem solving, verification, and *a priori* didactic analysis

Mathematical problem solving is a core component of lower secondary mathematics education and contributes to the development of mathematical thinking (Rendl et al., 2013). It is typically conceptualised as a multi-phase process that includes reflection and verification. Although the development of verification skills is critical to improving problem-solving ability, prior research in geometry problem solving suggests that learners often omit explicit verification, and when verification occurs it may remain cursory (Papadopoulos & Dagdilelis, 2008).

Understanding how problem solving processes unfold, and where typical breakdowns occur, is therefore an important concern for both researchers and teachers in lower secondary mathematics education.

The need for verification becomes particularly visible in tasks whose structure does not uniquely determine a single interpretation or solution pathway. Jonassen (1997) distinguishes well-structured problems from ill-structured problems, the latter requiring interpretative and evaluative judgements in addition to procedural work. In such tasks, a correct-looking sequence of steps may still lead to an inadequate solution if key assumptions or constraints remain unchecked. For teachers, this creates an instructional challenge: it is necessary not only to anticipate typical procedural errors, but also to foresee how students may interpret a task and where verification is likely to be missing. This anticipatory challenge connects directly to the practice of *a priori* didactic analysis in instructional planning.

Research on mathematical cognition shows that successful problem solving itself involves anticipatory thinking, such as foreseeing intermediate outcomes, control steps, and potential breakdowns during the solution process (Carlson & Bloom, 2005). From an instructional perspective, teachers need to anticipate these processes at the level of student thinking. This anticipatory dimension is captured by *a priori* task analysis, a concept rooted in the Theory of Didactical Situations (Brousseau, 1997) and operationalised for instructional planning by Nováková (2013), in which teachers consider likely student strategies, interpretations, and errors before classroom implementation.

However, systematic *a priori* analysis is cognitively demanding and often depends on teachers' experience and intuition, particularly in tasks requiring non-routine reasoning or explicit verification. This motivates the search for tools that could support teachers in anticipating student thinking during instructional planning. These open questions motivate the present study, which examines LLM outputs not in terms of correctness but with respect to process completeness and didactic relevance.

1.2 Large language models in mathematics education: opportunities and open questions

Recent literature highlights both opportunities and risks associated with the use of large language models (LLMs) in education. LLMs may support instruction by providing feedback, generating instructional materials, and assisting with individualisation (Kasneci et al., 2023). Reviews in mathematics education similarly suggest that tools such as ChatGPT can generate examples, explanations, and assessment items for both teachers and learners, while also raising concerns about superficial learning, academic integrity, and overreliance on AI-generated solutions (Almarashdi et al., 2024).

From a didactic perspective, however, important questions remain regarding LLMs as problem solvers and as generators of student-like solution attempts. Although LLMs can often produce correct final answers, it is less clear how they enact the underlying problem solving process and whether they spontaneously engage in essential phases such as verification. McGalliard and Otten (2025) examined AI responses to challenging mathematical problems and found that while models could handle routine tasks, their performance degraded on non-routine problems requiring insight or heuristic reasoning. This finding motivates empirical studies that examine LLM outputs not only in terms of correctness, but also with respect to process characteristics and didactic relevance.

1.3 Aim and research questions

The aim of this study is to explore the didactic potential of a large language model (LLM) for analysing mathematical problem solving processes and for supporting *a priori* didactic analysis in lower secondary mathematics education. Specifically, the study investigates how an LLM enacts key phases of mathematical problem solving under different prompting conditions and to what extent the same model can be used to generate student-like solution attempts that are meaningful for instructional planning.

The study is guided by three research questions that build on each other progressively. The first question establishes a baseline by examining the model's spontaneous problem solving behaviour. The second question tests whether structured prompting can improve the completeness of the process. The third question extends the investigation to the model's capacity for generating student-like solution attempts.

- *RQ1:* To what extent does an LLM, acting as an independent problem solver, enact the key phases of the mathematical problem solving process when solving lower secondary mathematics tasks?
- *RQ2:* How does the quality and completeness of the LLM's problem solving process change when the model is guided by a structured Pólya-inspired prompting framework?
- *RQ3:* To what extent can an LLM generate solution attempts that plausibly reflect different lower secondary student profiles in terms of ability and effort, and how useful are such outputs for *a priori* didactic analysis of typical errors, incomplete strategies, and instructional needs?

2 Theoretical background

2.1 Problem solving phases and verification as result checking

Problem solving in mathematics is commonly described as a process comprising several interrelated phases, including understanding the problem situation, devising a solution plan, carrying out this plan, and subsequently examining the obtained result. Although different authors use varying terminology, there is broad agreement that successful problem solving involves not only executing procedures but also assessing whether the produced result adequately satisfies the task conditions.

In his seminal work, Pólya (1945) conceptualises problem solving as a four-step process—understanding the problem, devising a plan, carrying out the plan, and looking back—where the final phase explicitly involves examining the result and checking the correctness of both the outcome and the procedure. Pólya's framework was adopted in this study for three reasons. First, it is the most widely recognised model of mathematical problem solving and provides a common reference point across the literature. Second, it explicitly identifies verification (looking back) as a distinct phase, which makes it well suited for operationalising verification as a coding criterion. Third, its four-step structure is naturally amenable to translation into a structured prompt for a language model. The use of this framework in the present study is pragmatic rather than theoretical: it serves as a scaffolding tool for the LLM, not as a theory of cognition.

Empirical studies indicate that the final phase is frequently omitted by novice solvers. In contrast to experts, who regularly monitor and verify their reasoning, beginners tend to apply familiar procedures quickly and uncritically, often without checking the adequacy of their solution (Schoenfeld, 1985). Kontorovich (2019) argues that students' failure to check solutions is not merely a matter of carelessness but reflects deeper epistemological dispositions about what constitutes a finished mathematical task.

In this study, the term *verification* is used in a narrow, operational sense to denote explicit checking of the correctness and plausibility of an obtained result. Such checking may include, for example, substituting a solution back into the original conditions, examining whether all constraints of the task have been respected, or providing a brief justification of why the result is valid. Verification in this sense does not necessarily involve extended metacognitive reflection but rather a concrete test of the solution's validity.

From a cognitive perspective, the frequent omission of verification can be interpreted through the lens of dual-process theories of reasoning. According to these theories, human thinking involves the interaction of two qualitatively different modes: a fast, intuitive, and heuristic mode (often referred to as System 1) and a slower, more deliberate, and analytical mode (System 2). While System 1 enables rapid generation of plausible solution steps, System 2 is typically required for careful monitoring, error detection, and explicit verification of results (Liu et al., 2025). In the absence of deliberate engagement of analytical control, problem solving may terminate prematurely once a seemingly acceptable result has been obtained. Liu et al. (2025) argues that contemporary LLMs tend to operate as black-box generators of fluent solution sequences, producing responses that resemble fast heuristic reasoning unless explicit mechanisms for verification are invoked. Their proposed *Invoke–Verify–Inject* framework highlights that verification does not emerge spontaneously but must be deliberately activated as a separate control process. This perspective provides theoretical support for interpreting verification as a fragile phase of problem solving that is easily omitted unless explicitly prompted.

At the same time, research on mathematical problem solving emphasises that verification should not be understood solely as a final, terminal step. Studies of open-ended and exploratory problem solving show that the process is inherently recursive rather than linear. Cifarelli and Cai (2005), for example, demonstrate that solvers continuously reformulate goals, pose subproblems, and adjust strategies while working on a task. Such recursive activity requires ongoing monitoring of intermediate results and assumptions, effectively embedding verification throughout the solution process rather than relegating it to the end.

From an educational perspective, this recursive view further explains why verification poses difficulties for novice solvers. Continuous monitoring and revision place high demands on cognitive control and metacognitive awareness, which are often underdeveloped in lower secondary students. As a result, verification may be reduced to an optional or entirely omitted activity, especially in the absence of explicit instructional emphasis. Treating verification as a distinct and observable component of problem solving therefore makes it possible to analyse whether this essential control mechanism is present or absent in a given solution, independently of the correctness of the final result.

By focusing on verification as explicit result checking, the present study adopts an operational perspective that is directly aligned with the evaluation criteria used in the empirical analysis. This approach enables a transparent examination of the completeness of problem solving processes while remaining consistent with the scope and aims of the study.

2.2 Ill-structured tasks and the need for interpretation

Mathematical tasks differ in the extent to which their structure constrains interpretation and solution pathways. This distinction is commonly captured through the concept of *ill-structured problems*, which are characterised by ambiguous conditions, incomplete specification of constraints, multiple reasonable assumptions, and the absence of a single uniquely defined solution (Jonassen, 1997). In contrast to well-structured tasks, ill-structured problems require solvers to make interpretative decisions about what the task situation means and which assumptions are admissible.

In mathematics education, ill-structuredness is often realised in classroom practice through *open-ended tasks*. Such tasks deliberately allow for multiple solution strategies, representations, or outcomes. While the terms *ill-structured* and *open-ended* are not strictly synonymous, they are closely related: open-ended tasks can be understood as a didactic instantiation of ill-structured problems, designed to foreground interpretation and evaluation as integral parts of mathematical activity.

In ill-structured and open-ended tasks, producing a procedurally correct result does not automatically guarantee that the task has been solved appropriately. The validity of a solution depends on whether the adopted interpretation, assumptions, and strategy are consistent with the task conditions. Explicit verification therefore becomes a central control mechanism, as it enables the solver to evaluate whether the obtained result is acceptable within the intended constraints and whether alternative interpretations would lead to different conclusions.

From an educational standpoint, open-ended tasks are also linked to the development of mathematical creativity. Empirical work suggests that open-ended problems can foster key components of creative mathematical thinking, including fluency, flexibility, and originality (Suyitno et al., 2018). At the same time, these tasks place substantially higher demands on self-monitoring and evaluation than routine problems. Without explicit instructional support, students may prioritise producing a plausible solution over critically examining whether it satisfies the conditions and assumptions embedded in the problem situation.

The importance of validation mechanisms in ill-structured tasks is highlighted within the Theory of Didactical Situations. Brousseau and Gibel (2005) argues that learning emerges through interaction between the learner and the problem environment (milieu), which provides feedback on the adequacy of proposed actions and interpretations. If the milieu does not offer meaningful feedback, learners lack resources for validating their implicit models and may remain unaware of inconsistencies in their reasoning. Under such conditions, verification is less likely to occur spontaneously because the need for it is not made visible by the task environment.

This perspective is directly relevant for analysing large language models in the role of problem solvers. When an LLM is presented with an ill-structured or open-ended task without external feedback and without explicit prompts for reflection, it operates in an environment analogous to a learner facing a non-responsive milieu. Under these conditions, the model—much like a student—may generate a fluent and procedurally coherent solution without recognising the need for explicit verification. This makes ill-structured and open-ended tasks a particularly sensitive context for examining verification as a fragile but didactically significant phase of the problem solving process.

2.3 *A priori* didactic analysis and anticipatory teacher knowledge

In mathematics education, *a priori* didactic analysis refers to systematic work with a task before it is implemented in the classroom, focusing on how students are likely to interpret the task, which strategies they may employ, and which errors or incomplete approaches may arise (Brousseau, 1997). In this sense, *a priori* analysis represents a form of anticipatory teacher knowledge that supports instructional planning by identifying critical points of reasoning, likely breakdowns, and opportunities for targeted intervention.

The ability to work productively with student errors is increasingly conceptualised as part of teachers' professional competence. Research on teacher noticing describes this competence as a process involving perception, interpretation, and decision-making. In the PID model proposed by Hoth et al. (2022), teachers must first notice a student's error, then interpret its underlying cause, and finally decide on an instructional response. While the PID model primarily describes in-the-moment classroom reasoning, *a priori* analysis can be seen as directly supporting its first two components by strengthening teachers' sensitivity to typical errors and their interpretative readiness.

Importantly, anticipatory accuracy is not fixed. Stannard and Foster (2025) provide evidence that secondary mathematics teachers can significantly improve their ability to predict student errors through repeated practice and comparison with empirical student data. However, such anticipatory work is time-demanding, especially for complex or non-routine tasks. Empirical findings suggest that expert teachers adjust their anticipation time based on task complexity; they devote substantially more time to anticipating typical student errors than novice teachers as task complexity increases, whereas they identify errors in simpler tasks more rapidly (Pankow et al., 2016). This highlights a practical tension: *a priori* analysis is didactically valuable, but often constrained by time and cognitive load.

From this perspective, large language models may serve as supportive tools for *a priori* didactic analysis by rapidly generating multiple plausible solution attempts, including incomplete reasoning, typical procedural errors, and missing verification steps. In the present study, such outputs are not interpreted as predictions of individual student behaviour, but as illustrative artefacts that may stimulate teachers' anticipatory thinking and support the interpretation phase of professional noticing.

At the same time, the use of AI-generated outputs in lesson preparation requires a critical stance. Recent work emphasises the need for *critical AI literacy* among educators, understood as the ability to evaluate AI outputs with awareness of their limitations, biases, and potential hallucinations (Ocak et al., 2025). When LLMs are used to simulate student thinking, their responses may appear pedagogically plausible while still misrepresenting actual student reasoning or overstating mathematical sophistication. Consequently, AI-generated solution attempts should be treated as material for reflection rather than as authoritative representations of students' cognitive processes.

2.4 Using large language models as simulated learners to support instructional planning

Recent research increasingly explores the use of artificial intelligence not only as a source of solutions or feedback, but also as a means of simulating learner behaviour for the purpose of teacher education and instructional planning. In this context, large language models have been proposed as *simulated learners*-artificial agents that generate plausible student-like solution attempts, including typical errors, misconceptions, and incomplete reasoning.

A key strength of LLM-based simulated learners is that they can generate student-like solution attempts that include not only procedural errors but also conceptually flawed interpretations of the task. In AI-driven role-play environments, teachers can therefore work with a range of plausible yet imperfect student responses and rehearse diagnostic reasoning and instructional moves in a low-risk setting (Zhuang & Zhang, 2025). Such simulations support a *decomposition of practice* by allowing teachers to focus on specific aspects of teaching (e.g., noticing and interpreting errors) without simultaneously managing the full complexity of classroom interaction (Zhuang & Zhang, 2025).

Zhuang and Zhang (2025) introduced *Student GPT*, a system designed to simulate erroneous mathematical reasoning in the domain of ratios, enabling pre-service teachers to practise diagnosing student misconceptions and selecting appropriate responses. Their findings suggest that repeated work with simulated student solutions allows teachers to practise and reflect on situations that are difficult to capture and revisit in real classrooms.

At the same time, empirical work cautions against assuming full fidelity of LLM-based student simulations. One limitation concerns the instability of role adherence: even when prompted to act as a struggling student, LLMs may shift into an expert or teacher-like register, which reduces the authenticity of the simulated learner and may distort the diagnostic task for the teacher (Zhuang & Zhang, 2025). More broadly, Guerra et al. (2025) argue that interactions with large language models can create an *illusion of understanding*, in which AI-generated responses appear coherent and pedagogically plausible while lacking grounding in human cognitive processes. From this perspective, simulated student solutions should not be interpreted as representations or predictions of actual student thinking, but rather as constructed artefacts whose plausibility depends on task type, prompting, and model characteristics.

Beyond student simulation, large language models have also been shown to support instructional preparation in the form of scaffolding and material design. Malik et al. (2025) demonstrate that, when appropriately prompted, LLMs can generate scaffolding sequences and preparatory tasks for middle school mathematics that are comparable in quality to those produced by expert educators. Such findings support the view that LLMs may contribute to *a priori* planning by helping teachers explore alternative task formulations, anticipate learning trajectories, and design targeted supports.

Taken together, these strands of research suggest that the educational value of large language models lies not in replacing teacher judgement, but in augmenting reflective instructional planning. When used as simulated learners or generators of preparatory materials, and when embedded within a framework of *a priori* analysis and critical interpretation, LLMs can support teachers in anticipating student thinking and identifying potential breakdowns in problem solving processes. From a cognitive perspective, these affordances and limitations are consistent with accounts that distinguish between rapid generation of plausible responses and slower processes of verification and integration, suggesting that AI-generated simulations should be treated as heuristic artefacts rather than models of fully grounded student cognition.

3 Methodology

3.1 Design overview

The study adopts a qualitative demonstration design aimed at exploring the problem solving behaviour of a large language model (LLM) and its potential didactic use in lower secondary mathematics education. Rather than evaluating model performance in terms of efficiency or optimisation, the focus is on analysing the structure and completeness of problem solving processes and on examining the didactic relevance of model-generated solution attempts.

The methodology comprises two complementary experiments. Experiment 1 investigates the behaviour of the LLM when solving lower secondary mathematics tasks under two different prompting conditions. In the baseline condition, the model acts as an independent problem solver without explicit methodological guidance. In the second condition, the model is guided by a structured Pólya-inspired framework that explicitly prompts different phases of problem solving. This experiment addresses Research Questions RQ1 and RQ2 by comparing the presence and quality of problem solving phases across conditions.

Experiment 2 explores the extent to which the same LLM can be used to generate student-like solution attempts by simulating different lower secondary student profiles. The aim of this experiment is not to predict individual student behaviour but to examine whether the generated outputs can plausibly reflect typical strategies and errors and whether they are useful for *a priori* didactic analysis. Experiment 2 addresses Research Question RQ3.

Across both experiments, the analysis is based on qualitative evaluation using explicitly defined criteria aligned with the theoretical framework outlined above. The methodological choices are designed to ensure transparency and consistency between the theoretical concepts, the experimental procedures, and the subsequent interpretation of results.

3.2 Model and experimental setting

All experiments were conducted using the Gemini 2.5 Pro large language model on 22 September 2025. The model was accessed in a standard interactive setting without any task-specific fine-tuning or additional training.

For each experimental condition, a single interaction thread was used. In Experiment 1, the full sequence of 24 tasks was presented within one interaction in the baseline condition and within a separate interaction in the Pólya-guided condition. In Experiment 2, separate interaction threads were used for each simulated student profile. No feedback was provided between tasks, and the model was not allowed to revise or correct its previous responses.

All tasks were presented to the model in Czech, and all responses were generated in Czech. The tasks were designed or selected to correspond to the mathematical level and forms of expression expected at the lower secondary level. Czech was used consistently as the language of interaction throughout the study. As Czech represents a less-resourced language in the training data of contemporary large language models, the observed problem solving behaviour may be influenced not only by task characteristics and prompting conditions but also by language-related factors.

In the baseline condition of Experiment 1, the model received no instructions regarding how to approach or present its solutions; the input consisted solely of the task statements. In the Pólya-guided condition and in Experiment 2, the model's behaviour was influenced only by the explicit prompts described in the corresponding methodological sections. No additional guidance regarding solution structure, verification, or expected length was provided beyond these prompts.

The experiments were conducted over a limited time period to reduce the potential impact of model updates. While the exact internal parameters of the model are not accessible, procedural transparency is ensured by clearly documenting the prompts, task set, and evaluation criteria used in the analysis.

3.3 Task set and task classification

The study is based on a set of 24 mathematical tasks selected for the purposes of the present analysis. The task set combines tasks designed or adapted by the authors with non-routine tasks. The non-routine tasks included in Appendix B were taken from publicly available problem sets of the Czech Mathematical Olympiad (MO). The tasks are reproduced for research and academic purposes with full source attribution. The original official wording and task context (year, round, category) are credited in the references. All tasks were chosen to be appropriate for the lower secondary level in terms of mathematical content, required procedures, and expected forms of reasoning.

To support a differentiated analysis of problem solving behaviour, the tasks were organised into four categories according to their structural characteristics and cognitive demands. The first category consisted of well-structured algebraic tasks, in which the problem conditions were explicitly stated and a standard solution procedure could be applied. These tasks served as a reference point for analysing basic procedural correctness and explicit result checking.

The second category comprised ill-structured or open-ended word problems designed or adapted by the authors. These tasks allowed for ambiguity in interpretation or admitted multiple reasonable solution approaches and required solvers to make interpretative decisions. They provided a suitable context for examining whether explicit verification was used to check the adequacy of the adopted interpretation.

The third category included non-routine problems taken directly from the Czech Mathematical Olympiad, category Z9, which targets mathematically advanced ninth-grade students (aged 14–15). They typically require insight, heuristic reasoning, or a non-standard strategy and are substantially more demanding than routine classroom tasks. Research has shown that AI models may struggle with such non-routine problems even when performing well on standard tasks (McGalliard & Otten, 2025). The inclusion of these tasks therefore provides a stringent test of the model's problem solving behaviour in mathematically demanding situations.

The fourth category consisted of geometric construction tasks designed by the authors. These tasks required solvers to propose and justify a construction procedure and to consider the validity of the resulting construction.

Each category contained six tasks, resulting in a total of 24 tasks. The same task set was used in both prompting conditions in Experiment 1. In Experiment 2, a subset of tasks was selected from each category in order to reduce the overall number of simulated solutions while preserving the diversity of task types.

3.4 Experiment 1: LLM as problem solver

Experiment 1 examined how the large language model responded to lower secondary mathematics tasks under two different conditions: (a) without any instructional prompting beyond the task statements themselves, and (b) with an explicit, structured prompt inspired by Pólya's problem solving framework.

3.4.1 Procedure and conditions

In both prompting conditions, the model interacted with the full set of 24 tasks within a single interaction thread. The tasks were presented sequentially, one after another, without feedback or correction between tasks.

In the baseline condition, the model received no instructions regarding how to approach or present its solutions. The input consisted solely of the task statements provided in Czech. No guidance was given concerning solution structure, explanation of reasoning, or result checking. This condition was intended to capture the model's spontaneous responses to a sequence of mathematical tasks.

In the Pólya-guided condition, a new interaction thread was initiated. At the beginning of the interaction, the model received a structured introductory prompt inspired by Pólya's problem solving framework. This prompt instructed the model to address four phases of problem solving: understanding the problem, devising a plan, carrying out the plan, and looking back, with explicit emphasis on result checking in the final phase. The full wording of this prompt, in its original Czech form, is provided in Appendix A. The same 24 tasks were then presented sequentially within the same interaction thread. After each task statement, a brief reminder was included prompting the model to adhere to the problem solving framework introduced at the beginning of the interaction. No feedback was provided between tasks, and the model was not allowed to revise or correct its previous responses. The full wording of the 24 tasks (in Czech) is provided in Appendix B. The complete transcripts of all model interactions, including all task statements and verbatim model-generated responses for both experiments (approx. 200 pages), are provided as Supplementary Material and are available at: https://https://osf.io/k2vpw/.

3.4.2 Evaluation criteria for Experiment 1

The solutions generated in Experiment 1 were analysed using a set of qualitative evaluation criteria focusing on the presence and completeness of key phases of the problem solving process. The analysis did not aim to assess mathematical elegance or efficiency but rather to examine whether essential components of a complete problem solving process were present in the model's responses.

Each solution was evaluated independently with respect to four criteria: (1) understanding of the task, (2) solution process, (3) correctness of the result, and (4) verification as result checking. The criteria were applied consistently across both prompting conditions.

Understanding of the task This criterion captured whether the model's response demonstrated an appropriate interpretation of the task. A solution was considered to show understanding if the subsequent reasoning addressed the relevant task conditions and goals. Misinterpretations of the task statement or solutions addressing a different problem were coded as lacking task understanding, regardless of the correctness of intermediate steps.

Solution process The solution process criterion focused on whether the model provided a coherent sequence of steps leading from the interpretation of the task to a proposed result. Solutions were coded positively if they included a logically connected procedure or line of reasoning. Fragmented, incomplete, or internally inconsistent solution attempts were coded negatively, even if a numerical result was present.

Correctness of the result Correctness referred to whether the final result produced by the model was mathematically correct with respect to the adopted interpretation of the task. This criterion was evaluated independently of the presence or absence of explicit verification. A solution could therefore be correct without being verified, or verified without being correct.

Verification as result checking Verification was operationalised as the explicit checking of the obtained result with respect to the task conditions. This included, for example, substituting the result back into the original conditions, explicitly confirming that all constraints had been satisfied, or providing a brief justification of why the result was valid. Implicit confidence statements or unexamined final answers were not considered evidence of verification.

For each criterion, solutions were coded using a three-level scale (0–2), where 0 indicated absence, 1 partial or ambiguous presence, and 2 clear and explicit presence of the respective feature. The coding scheme was applied to all solutions generated in Experiment 1 and formed the basis for the subsequent qualitative comparison between the baseline and the Pólya-guided conditions.

3.5 Experiment 2: LLM as simulated learner

The experiment was conducted using the same model and interaction setting as in Experiment 1. All prompts and task statements were provided in Czech, and all responses were generated in Czech.

3.5.1 Student profiles

Three student profiles were defined to reflect differences commonly observed among lower secondary students with respect to mathematical proficiency, persistence, and attention to result checking. The profiles were informed by typologies of student performance and typical difficulties described in the literature, in particular by the classification of student solution types discussed by Rendl et al. (2013).

Each profile was described in a short but explicit narrative specifying typical characteristics of the student's approach to mathematical problem solving. The profiles were intentionally formulated in a non-diagnostic manner and were not intended to represent real individual students. Their purpose was to provide a didactically meaningful contrast between different types of solution attempts that teachers may anticipate when planning instruction. The full wording of the student profile prompts, in their original Czech form, is provided in Appendix C.

3.5.2 Procedure and task selection

From each of the four task categories described earlier, two tasks were selected, resulting in a total of eight tasks used in Experiment 2. This selection was made to preserve task diversity while limiting the overall number of generated solutions.

For each student profile, a separate interaction thread was initiated. At the beginning of each thread, the model received a detailed description of the corresponding student profile. The selected tasks were then presented sequentially within the same thread. This design was chosen to maintain consistency in the simulated student behaviour across multiple tasks.

As in Experiment 1, no feedback was provided between tasks, and the model was not instructed to revise or correct its solutions. In contrast to the baseline condition of Experiment 1, the prompts in Experiment 2 explicitly framed the model's role as that of a student rather than an independent problem solver.

3.5.3 Evaluation criteria for Experiment 2

The generated solution attempts were analysed qualitatively with respect to role fidelity and didactic usefulness.

Role fidelity Role fidelity referred to the extent to which the generated solution attempts were consistent with the specified student profile across tasks. Solutions were considered to show high role fidelity if the level of mathematical reasoning, types of errors, degree of persistence, and presence or absence of result checking aligned with the profile description throughout the interaction thread.

Didactic usefulness Didactic usefulness captured whether the generated solution attempts could support teachers in *a priori* didactic analysis. Solutions were considered didactically useful if they illustrated typical student errors, incomplete reasoning, missing or incorrect verification steps, or partially successful strategies that could reasonably occur in classroom practice. Such solution attempts may help teachers anticipate critical moments in instruction, prepare targeted questions, or plan follow-up activities. Fully polished expert-like solutions or implausible student responses were considered less useful from a didactic perspective.

The analysis in Experiment 2 was exploratory and qualitative in nature and aimed to identify illustrative patterns rather than to quantify frequencies or make predictive claims.

3.6 Trustworthiness of the design

Several design choices support the trustworthiness of the study. The experimental procedures, prompts, and task set are documented in detail, including the original Czech wording of all prompts and tasks (Appendices A–C). The use of the same task set across conditions in Experiment 1 enables a focused comparison of problem solving behaviour with and without structured guidance. Finally, the explicit evaluation criteria provide a transparent basis for the qualitative coding.

4 Results

4.1 Experiment 1: Completeness of problem solving phases and result verification

4.1.1 Overall comparison of baseline and Pólya-guided conditions

Table 1 reports the number of tasks in which selected problem solving components were present in each experimental condition. A component was considered present if it was coded as partially or clearly present (codes 1 or 2).

As shown in the table, the model produced a response that included explicit task understanding, a coherent solution process, and a final result for all tasks in both conditions. These components were therefore present across the entire task set.

In contrast, clear differences between conditions were observed with respect to explicit result verification. In the baseline condition, explicit verification was observed in only 7 out of 24 tasks, whereas in the Pólya-guided condition verification was present in all 24 tasks.

Table 1: Occurrence of key problem solving components across tasks in Experiment 1 ($n = 24$)

Problem solving component	*Baseline*	*Pólya-guided*
Explicit task understanding	24	24
Coherent solution process	24	24
Correct final result	24	24
Explicit result verification	7	24

4.1.2 Verification across task categories in the baseline condition

Table 2 shows the occurrence of explicit result verification across different task categories in the baseline condition. Verification was unevenly distributed across task types.

Table 2: Occurrence of explicit result verification across task categories in the baseline condition (Experiment 1, $n = 24$)

Task category	*Tasks with verification*	*Total tasks*
Category A: Well-structured algebraic tasks	4	6
Category B: Ill-structured word problems	1	6
Category C: Non-routine (olympiad) problems	2	6
Category D: Geometric construction tasks	0	6

Figure 1 presents these results graphically, making the contrast between the two conditions immediately visible across all task categories.

Explicit verification occurred most frequently in well-structured algebraic tasks, where it was present in four out of six tasks. In ill-structured word problems, verification was observed only once. In non-routine olympiad problems, explicit verification occurred in two tasks. No explicit verification was observed in any of the geometric construction tasks.

These results indicate that, in the absence of explicit guidance, the occurrence of verification depended on task characteristics.

In the Pólya-guided condition, explicit result verification was present in all tasks across all task categories.

4.1.3 Illustrative examples of verification behaviour

To provide a richer picture of the observed patterns, we present selected examples illustrating how verification was enacted or omitted across task categories.

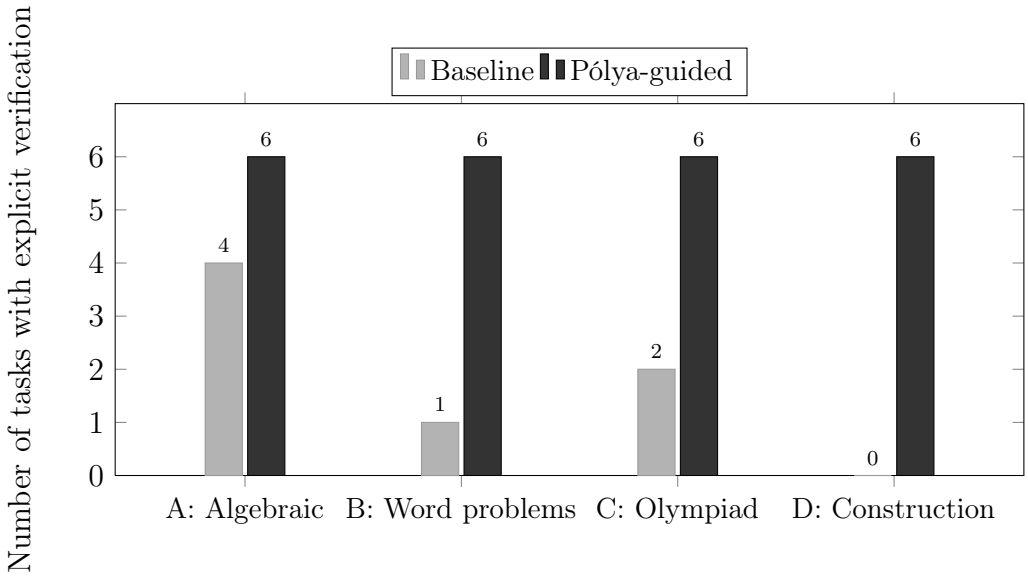

Fig. 1: Occurrence of explicit result verification across task categories in the baseline and Pólya-guided conditions (Experiment 1, $n = 6$ tasks per category)

Verification present (Task A1, baseline) For the equation $3x - 7 = 2x + 5$, the model produced a clean solution with all intermediate steps and concluded with an explicit substitution check: it substituted $x = 12$ back into both sides of the equation and confirmed equality. This exemplifies verification as explicit result checking in its most straightforward form.

Verification absent (Task B1, baseline) For the word problem about a car travelling at 60 km/h for 2 hours including a break, the model correctly identified the ambiguity (the break duration is unknown, so the distance cannot be uniquely determined) and presented alternative cases. However, no verification was performed: the model did not check whether its proposed cases were consistent with the problem conditions, nor did it reflect on whether additional interpretations existed.

Verification absent (Task D4, baseline) For the construction of a rhombus with given diagonals, the model described a correct and elegant construction procedure using the perpendicular bisector. However, no final verification was provided—the model did not check that the constructed quadrilateral indeed possessed all required properties of a rhombus (equal sides, perpendicular bisecting diagonals).

Verification present (Task D4, Pólya-guided) The same task, when solved under the Pólya-guided condition, included an explicit verification step. After completing the construction, the model confirmed that the diagonals bisect each other perpendicularly and that all four resulting right triangles are congruent, thereby establishing that the figure is indeed a rhombus.

Detailed task-level coding results and qualitative notes supporting the evaluation for both experimental conditions are provided in Appendices D and E.

4.2 Experiment 2: Simulation of student solution profiles

Experiment 2 examined whether the large language model could generate solution attempts that plausibly correspond to different lower secondary student profiles and whether such solutions are didactically useful. The evaluation did not focus on mathematical correctness but on (a) alignment with the intended student profile and (b) didactic usefulness of the generated solutions.

Three student profiles were considered: a diligent but low-achieving student, a mathematically able but unmotivated student, and a student with neither strong ability nor motivation. For each profile, a subset of eight tasks was analysed. Figures 2–4 summarise the results for the three profiles using a three-level scale (0–2). Values of 0, 1, and 2 indicate low, partial, and high presence of the respective criterion (role alignment or didactic usefulness).

Across the three profiles, the generated solution attempts differed in both role alignment and didactic usefulness. Figures 2–4 present these results graphically. In routine lower secondary tasks from categories A and B, the model often produced solution attempts that matched the intended profiles and contained

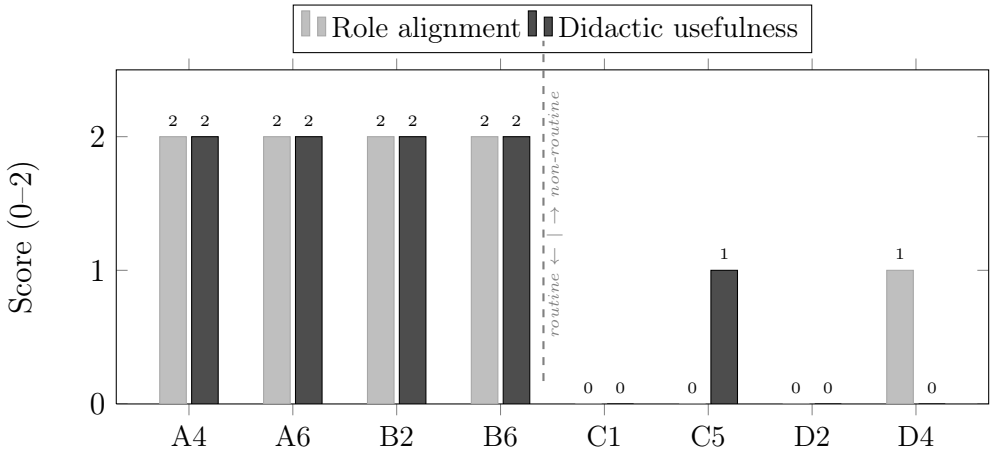

Fig. 2: Experiment 2 — Diligent but low-achieving student: role alignment and didactic usefulness across tasks (0–2 scale). The dashed line separates routine tasks (Categories A, B) from non-routine and construction tasks (Categories C, D)

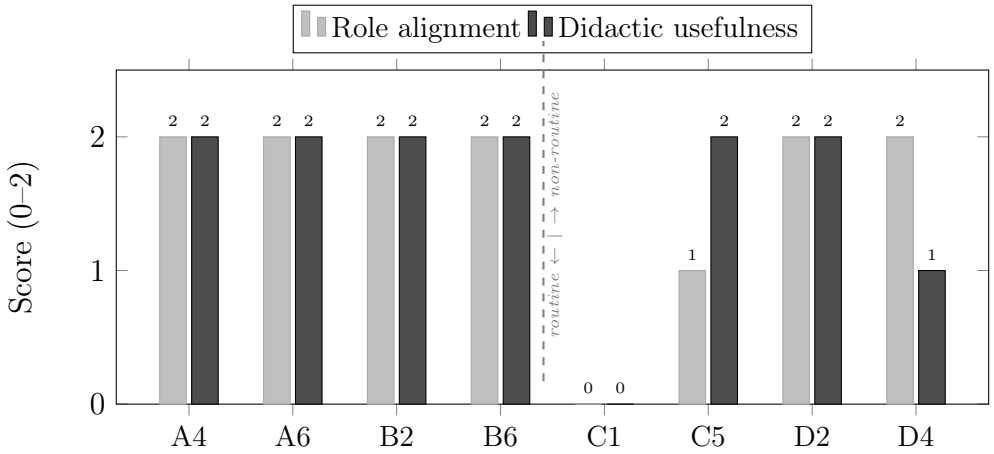

Fig. 3: Experiment 2 — Able but unmotivated student: role alignment and didactic usefulness across tasks (0–2 scale). The dashed line separates routine tasks from non-routine and construction tasks

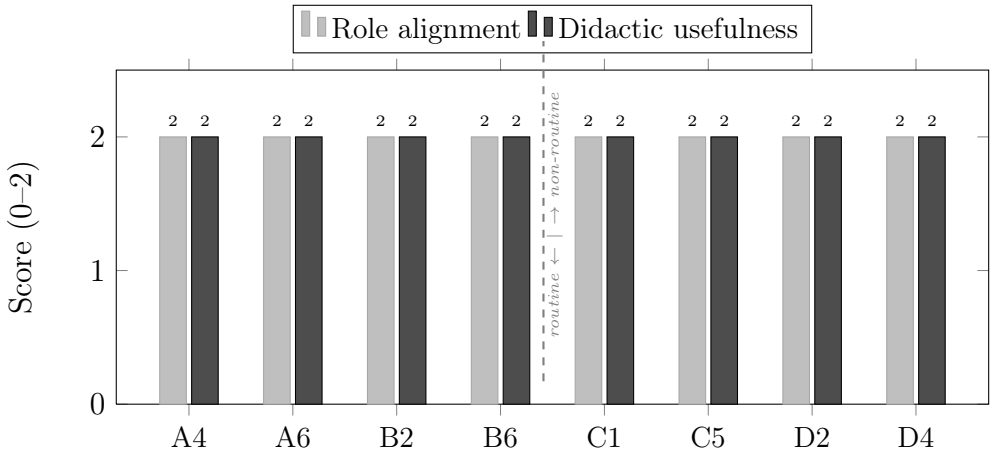

Fig. 4: Experiment 2 — Struggling student: role alignment and didactic usefulness across tasks (0–2 scale). Unlike the other two profiles, both criteria remain at the maximum level across all tasks, including non-routine and construction tasks

plausible student-like errors or omissions. Such errors were treated as didactically valuable, as they can serve as material for *a priori* analysis (e.g., anticipating typical pitfalls and planning teacher prompts).

However, the results also show clear limits of the simulation approach. For the diligent but low-achieving profile, role alignment deteriorated in several non-routine and construction tasks (categories C

and D), where the model sometimes produced solutions that were too advanced or too systematically argued to be plausible for the intended student type. A similar tendency was observed for the able-but-unmotivated profile in the most demanding tasks, although some responses still reflected the intended profile through shortcuts, missing checks, or inconsistent reasoning.

In contrast, the struggling student profile showed consistently high role alignment across all selected tasks, with both criteria reaching the maximum score of 2 for every task including the non-routine and construction categories. The corresponding solution attempts frequently included incomplete reasoning, breakdowns in procedure, or missing result checking, which were considered didactically useful as representations of typical difficulties that teachers may encounter. This uniform pattern stands in marked contrast to the other two profiles and confirms that the model more reliably simulates lower levels of performance than higher-but-constrained levels.

Detailed qualitative analyses of individual solution attempts for all three profiles are provided in Appendix F.

4.2.1 Qualitative illustrations of role alignment and didactic usefulness

The following vignettes illustrate both successful and unsuccessful student simulations across profiles and task types.

Successful simulation—diligent but low-achieving student, Task A4 The model produced a solution attempt in which the student correctly identified a common denominator but made a typical sign error when removing parentheses (failing to distribute the minus sign). The condition $x \neq 1$ was omitted entirely. The language was diligent but mechanical, consistent with the profile. This type of error—a minus sign in front of parentheses—is a well-documented difficulty at the lower secondary level, making this output directly useful for a priori analysis.

Successful simulation—able but unmotivated student, Task B2 The model simulated a student who quickly added the two discounts ($20\% + 10\% = 30\%$) and treated the discounted price as 70% of the original, ignoring the fact that percentage discounts are applied sequentially and multiplicatively. The resulting answer was incorrect, but the error pattern—additive composition of percentages—is a frequent and well-documented student misconception. The output clearly matched the intended profile (fast, intuitive, no verification) and provided useful diagnostic material.

Failed simulation—diligent but low-achieving student, Task C1 For this olympiad problem requiring prime factorisation and comparison of exponents, the model produced a systematic, complete, and error-free solution that would be unrealistic even for mathematically gifted ninth-grade students, let alone for the intended profile of a diligent but untalented student. The solution employed parametrisations and systematic minimisation procedures characteristic of university-level work. This output entirely missed the intended student profile and offered no didactic value for anticipating the difficulties of the targeted student type.

Consistently successful—struggling student, all tasks The struggling student profile showed the highest role alignment across all tasks. For Task D4 (rhombus construction), the model produced a superficial attempt without analysis or verification, with a key error of confusing the diagonal's length with its half. For Task C1 (olympiad problem), the model responded with a simple numerical substitution attempt followed by resignation. These outputs plausibly represent the kinds of difficulties teachers encounter with struggling students and can support anticipatory planning.

5 Discussion

The present study examined the didactic potential of a large language model for analysing mathematical problem solving processes and supporting *a priori* didactic analysis in lower secondary mathematics education. The focus was deliberately not on mathematical performance or optimisation, but exclusively on the structure, completeness, and didactic relevance of generated solution attempts.

The present findings are consistent with previous studies emphasising the supportive yet limited role of LLMs in mathematics education. Similar to the observations of Pepin et al. (2025), our results suggest that LLMs can assist with instructional planning and reflection. In line with Kasneci et al. (2023), the study also reinforces concerns that uncritical reliance on AI-generated solutions may reduce engagement in verification and deeper mathematical reasoning.

5.1 Verification as a fragile phase of problem solving

Results from Experiment 1 show that explicit result verification represents a particularly fragile phase of the problem solving process. In the baseline condition, verification was frequently omitted despite the presence of correct interpretations, coherent solution procedures, and correct final results. This finding is consistent with long-standing observations in mathematics education that students often stop once a result is obtained and do not spontaneously engage in checking or reflection (Kontorovich, 2019).

The absence of verification was not evenly distributed across task types. It occurred most frequently in ill-structured word problems and geometric construction tasks, where correctness cannot be reduced to a single numerical outcome. These task types require solvers to reflect on assumptions, constraints, and the adequacy of the chosen interpretation. The results suggest that, without explicit prompting, even a highly capable language model tends to prioritise producing a plausible solution over critically examining its validity.

When guided by a structured Pólya-inspired framework, explicit verification was present in all tasks across all categories. We acknowledge that this finding may appear predictable: if the prompt explicitly requests verification, it is not surprising that the model produces it. However, the empirical demonstration of this contrast is valuable for two reasons. First, it quantifies the gap between prompted and unprompted conditions across different task types, which has not been documented for lower secondary mathematics tasks in Czech. Second, it has direct practical implications for teachers: structured prompting is not merely a theoretical recommendation but a necessary condition for obtaining didactically complete outputs.

This sensitivity to prompting is consistent with recent findings on how users elicit higher-quality reasoning from LLMs. Urban et al. (2025) report that successful users of ChatGPT often include explicit evaluation criteria in their prompts. Our results suggest a similar mechanism: the Pólya-inspired prompt did not change the model's mathematical ability, but it consistently elicited verification, which was frequently missing in the baseline condition. This indicates that structured prompting can serve as a practical scaffolding tool for activating checking behaviour.

5.2 LLMs as tools for analysing problem solving processes

The findings of Experiment 1 suggest that large language models can be used as analytical tools for examining problem solving processes. The baseline condition revealed a systematic pattern: the model consistently prioritised procedural completion over explicit reflection or result checking. As the illustrative examples in Section 4.1.3 show, even when the model identified ambiguity (as in Task B1) or produced elegant constructions (as in Task D4), it did not spontaneously engage in verification.

At the same time, the results underscore the importance of prompt design. Structured prompts did not merely change the form of the output but fundamentally altered the completeness of the problem solving process. This sensitivity to prompting highlights both the potential and the limitation of LLMs: they can support didactic analysis, but they do not autonomously prioritise pedagogically desirable behaviours such as verification or reflection.

5.3 Simulating student thinking: Didactic potential and limits

Results from Experiment 2 show that, for routine and moderately demanding tasks (Categories A and B), the model often produced profile-consistent solution attempts containing plausible student-like errors. The diligent student's sign error in Task A4, the gifted student's additive percentage misconception in Task B2, and the struggling student's procedural breakdown in Task B6 all represent documented patterns that teachers may encounter in classroom practice. From a didactic perspective, such imperfect solutions are valuable because they make typical breakdowns in problem solving processes visible and can therefore serve as concrete material for anticipatory instructional planning.

This aligns with research on LLM-based simulated learners, which suggests that simulated student responses can support practice-based work with student errors and misconceptions (Zhuang & Zhang, 2025). The present study extends this perspective by showing that the plausibility of such simulations is strongly task-dependent.

However, the results also demonstrate clear limits. For non-routine and construction tasks, the model frequently produced solutions that were too advanced for the intended profiles. This was most striking for the diligent but low-achieving profile: in Tasks C1 and D2, the model generated solutions at a level of sophistication that would be unrealistic even for mathematically gifted students (see Section 4.2.1). This deterioration of role fidelity occurred precisely in the task categories where anticipatory teacher knowledge is most needed, consistent with the finding of McGalliard and Otten (2025) that AI performance patterns differ between routine and non-routine problems.

The struggling student profile was the exception: it showed consistently high role alignment across all tasks, suggesting that the model more reliably simulates lower levels of performance than higher-but-constrained levels. This asymmetry has practical implications for the use of LLM-generated solutions in a priori analysis.

These findings are consistent with broader limitations identified in the literature. Zhuang and Zhang (2025) observe that simulated learners often adapt unrealistically quickly to instructional input and may respond in a compliant and passive manner, without the hesitation, partial attempts, or follow-up questions that characterise genuine student behaviour. Such limitations restrict the extent to which LLM-based simulations can approximate the dialogical dynamics of classroom diagnosis.

Role fidelity is also a matter of linguistic register. Even when instructed to adopt the voice of a lower secondary student, LLMs may produce language that is overly formal or "teacher-like" (Zhuang & Zhang, 2025). Pando and León (2025) similarly found that achieving age-appropriate language required iterative prompting with explicit output constraints, rather than role prompts alone. This is consistent with our observation that simulated profiles sometimes drifted toward expert-like discourse.

More broadly, LLM-generated responses may exhibit what Guerra et al. (2025) term an *illusion of understanding*: outputs that appear coherent while lacking deeper conceptual grounding. In the context of student simulation, this may lead to overestimating the cognitive realism of generated solutions. One contributing factor is that student profiles are specified only narratively, leaving room for the model to improvise competence. Zhang et al. (2025) suggest that structured representations of learner proficiency (e.g., a "Skill-Tree") can help constrain such drift, which may be a promising direction for improving simulation fidelity in non-routine tasks.

Finally, logical inconsistency across tasks remains a concern. In our data, the diligent but low-achieving profile alternated between elementary errors in routine tasks and unexpectedly sophisticated reasoning in olympiad problems (compare Tasks A4 and C1 in Figure 2). While some degree of inconsistency may resemble authentic student behaviour, the abrupt shifts observed here exceeded what would be plausible for a single student and require cautious interpretation by the teacher.

5.4 Implications for *a priori* didactic analysis

Taken together, the results of both experiments suggest a specific and circumscribed didactic role for large language models in mathematics education: supporting *a priori* didactic analysis by making potential solution paths, typical errors, and missing problem solving phases visible before instruction. In this role, the value of LLM-generated solutions lies not in their correctness or realism, but in their capacity to stimulate teachers' anticipatory thinking.

Building on research showing that error anticipation can be developed through practice (Stannard & Foster, 2025), the present findings suggest that LLM-generated solutions may serve as a readily available source of varied examples for such anticipatory work. In this sense, LLMs can reduce the initial time and effort required to generate and compare alternative student-like solution attempts during lesson preparation. At the same time, the limitations of simulation highlight the need for critical AI literacy: teachers should treat AI-generated outputs as prompts for professional judgement rather than as proxies for real student thinking.

The present findings further underscore the instructional importance of explicit result verification. If even a language model with strong procedural capabilities omits verification unless explicitly prompted, this reinforces the need to foreground result checking as an explicit instructional goal, particularly in ill-structured and open-ended tasks (Schoenfeld, 1985).

These implications also reinforce the importance of critical AI literacy among teachers. In their systematic review of ChatGPT in school mathematics education, Turmuzi et al. (2026) emphasise that AI integration requires teachers to remain the primary facilitators who critically evaluate and contextualise AI outputs. Our results provide a concrete illustration of why this is necessary: the model not only drifted into unrealistic student simulations in cognitively demanding tasks (e.g., producing expert-like constructions for a low-achieving profile), but also tended to omit explicit result verification unless it was systematically prompted. Uncritical adoption of such outputs could therefore lead teachers to anticipate incorrect difficulties, overlook missing checking behaviour, and design misguided instructional responses. Consequently, LLM-generated solution attempts should be treated as reflective artefacts for professional judgement rather than as reliable proxies for authentic student thinking.

6 Methodological limitations

The present study was designed as a qualitative demonstration study, and its findings should be interpreted with this scope in mind. The aim was not to establish generalisable claims about the performance

of large language models but to provide a transparent and analytically grounded examination of their problem solving behaviour under specific conditions relevant to mathematics education.

Several limitations of the design need to be acknowledged. First, the experiments were conducted using a single large language model and a single model version. The observed behaviour may therefore reflect characteristics of this particular model rather than properties shared by all contemporary LLMs.

Second, all interactions were conducted in Czech, which represents a less-resourced language in the training data of most LLMs. Language-related factors may thus have influenced the completeness and structure of the generated solutions.

Third, the analysis relied on qualitative coding performed by the authors. While explicit evaluation criteria were used to support consistency, no claims are made regarding inter-rater reliability. The purpose of the analysis was exploratory and illustrative rather than statistical. Similarly, the student profiles used in Experiment 2 were intentionally simplified and do not correspond to diagnostic categories or individual learners.

Fourth, the empirical basis consists of outputs from a single model and a limited number of generated responses. While LLM outputs could in principle be produced in larger quantities, the qualitative demonstration design required detailed multi-criteria analysis of each solution, which limited the feasible sample size. Additionally, simulations conducted through text-only interfaces necessarily omit non-verbal and multimodal information that teachers routinely use when interpreting student difficulties (Guerra et al., 2025).

Fifth, the development of AI technologies is extremely rapid, and the findings reflect the behaviour of a specific model version (Gemini 2.5 Pro, September 2025). Future model updates may alter the observed patterns. However, the methodological framework and evaluation criteria developed in this study remain applicable to future models, which is one of the intended contributions of the paper.

7 Future research

Future research could extend this work by systematically comparing different models and languages, involving independent coders, and examining how teachers interpret and use LLM-generated solutions in authentic lesson planning contexts. Multi-model comparative studies using larger samples with streamlined coding procedures would help establish the generalisability of the observed patterns. A particularly important direction is the development of multimodal and interactionally richer simulated learners that better approximate classroom diagnosis beyond text-only outputs. Future work could also explore whether structured learner models (e.g., Skill-Tree representations) improve the stability of student simulations, particularly in non-routine and construction tasks. Another promising direction is to test iterative prompting strategies that explicitly constrain linguistic register and mathematical tool use to match a targeted age group.

8 Conclusion

This study explored the didactic potential of a large language model as a tool for analysing mathematical problem solving processes and supporting *a priori* didactic analysis in lower secondary mathematics education. The focus was deliberately placed on the structure and completeness of solution processes and on their didactic relevance, rather than on mathematical performance or optimisation.

Experiment 1 showed that explicit result verification represents a fragile phase of problem solving that is often omitted unless it is explicitly prompted. The contrast between the baseline and the Pólya-guided conditions suggests that the absence of verification should not be interpreted as a lack of capability but rather as a tendency not to activate this phase spontaneously. From a didactic perspective, this finding reinforces the importance of explicitly foregrounding verification and reflection in instructional design, particularly in ill-structured tasks and construction problems.

Experiment 2 demonstrated that a large language model can generate solution attempts that plausibly reflect certain student-like behaviours, including typical errors, omissions, and incomplete strategies. At the same time, clear limits emerged. For non-routine and olympiad-type problems, the model tended to produce highly sophisticated or expert-level solutions even when prompted to simulate an average or low-achieving student. As a result, profile alignment deteriorated precisely in tasks where anticipatory teacher knowledge is most needed. These findings suggest that LLM-based student simulation is strongly task-sensitive: it can support anticipatory work for routine tasks, but its use for mathematically demanding problems requires particular caution.

Overall, the study suggests that large language models may serve as reflective tools for didactic analysis when used critically and transparently. Their value lies not in replacing teacher expertise or

evaluating student performance, but in making aspects of problem solving visible and open to reflection. Further research is needed to examine how such tools can be integrated into instructional planning in a responsible and pedagogically meaningful way.

Declarations

Funding: The authors received no specific funding for this work.
Conflict of interest: The authors declare no conflict of interest.
Generative AI use: The authors used ChatGPT to support English-language editing and improve clarity of phrasing. All scientific content, study design, analysis, interpretation, and final wording were reviewed and approved by the authors.

References

Almarashdi, H. S., Jarrah, A. M., Abu Khurma, O., & Gningue, S. M. (2024). Unveiling the potential: A systematic review of ChatGPT in transforming mathematics teaching and learning. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(12), em2555. https://doi.org/10.29333/ejmste/15739

Brousseau, G. (1997). *Theory of Didactical Situations in Mathematics*. N. Balacheff, M. Cooper, R. Sutherland, & V. Warfield (Eds.) Kluwer Academic Publishers.

Brousseau, G., & Gibel, P. (2005). Didactical handling of students' reasoning processes in problem solving situations. *Educational Studies in Mathematics*, *59*(1–3), 13–58. https://doi.org/10.1007/s10649-005-2532-y

Carlson, M. P., & Bloom, I. (2005). The cyclic nature of problem solving: An emergent multidimensional problem-solving framework. *Educational Studies in Mathematics*, *58*(1), 45–75. https://doi.org/10.1007/s10649-005-0808-x

Cifarelli, V. V., & Cai, J. (2005). The evolution of mathematical explorations in open-ended problem-solving situations. *Journal of Mathematical Behavior*, *24*(3–4), 302–324. https://doi.org/10.1016/j.jmathb.2005.09.007

Guerra, E., Peña, M., & Araya, R. (2025). The inevitable and unpredictable role of large language models in education: A commentary on Huettig and Christiansen (2024). *Cognitive Science*, *49*, e70105. https://doi.org/10.1111/cogs.70105

Hoth, J., Larrain, M., & Kaiser, G. (2022). Identifying and dealing with student errors in the mathematics classroom: Cognitive and motivational requirements. *Frontiers in Psychology*, *13*, 1057730. https://doi.org/10.3389/fpsyg.2022.1057730

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65–94. https://doi.org/10.1007/BF02299613

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kontorovich, I. (2019). Why do students not check their solutions to mathematical problems? A field-based hypothesis on epistemological status. *International Journal of Mathematical Education in Science and Technology*, *50*(7), 1050–1062. https://doi.org/10.1080/0020739X.2019.1650304

Liu, J., Huang, Z., Liu, Q., Ma, Z., Zhai, C., & Chen, E. (2025). Knowledge-centered dual-process reasoning for math word problems with large language models. *IEEE Transactions on Knowledge and Data Engineering*, *37*(6), 3457–3471. https://doi.org/10.1109/TKDE.2025.3556367

Malik, R., Abdi, D., Wang, R., & Demszky, D. (2025). Scaffolding middle school mathematics curricula with large language models. *British Journal of Educational Technology*, *56*(3), 999–1027. https://doi.org/10.1111/bjet.13571

Matematická olympiáda. (2021). Czech Mathematical Olympiad: 71st year, category Z9, Round I (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3492117/z71i-9.pdf

Matematická olympiáda. (2022a). Czech Mathematical Olympiad: 72nd year, category Z9, Round I (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3494370/z72i-9.pdf

Matematická olympiáda. (2022b). Czech Mathematical Olympiad: 72nd year, category Z9, Round II (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3516462/z72ii-9-zr.pdf

Matematická olympiáda. (2022c). Czech Mathematical Olympiad: 72nd year, category Z9, Round III (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3516464/z72iii-9-zr.pdf

Matematická olympiáda. (2024a). Czech Mathematical Olympiad: 74th year, category Z9, Round II (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3828323/z74ii-zadani.pdf

Matematická olympiáda. (2024b). Czech Mathematical Olympiad: 74th year, category Z9, Round III (problem set). Retrieved February 8, 2026, from https://www.matematickaolympiada.cz/media/3824271/z74iii-9.pdf

McGalliard, W., & Otten, S. (2025). AI Responses to Challenging Problems and Educator Responses to AI Availability. *Digital Experiences in Mathematics Education*, *11*, 319–332. https://doi.org/10.1007/s40751-024-00167-4

Nováková, H. (2013). Analýza a priori jako součást přípravy učitele na výuku. *Scientia in educatione*, *4*(2), 20–51. https://doi.org/10.14712/18047106.70

Ocak, C., Kopcha, T. J., Hodges, C. B., Sadik, O., & Ozogul, G. (2025). How artificial intelligence will reshape education: Conversations with the educational technology community. *TechTrends*, *70*, 330–345. https://doi.org/10.1007/s11528-025-01149-y

Pando, M., & León, M. (2025). Mathematics disciplinary literacy: A case study of a bilingual teacher's interaction with ChatGPT. *Language and Education*, *40*(4), 980–999. https://doi.org/10.1080/09500782.2025.2601055

Pankow, L., Kaiser, G., Busse, A., König, J., Blömeke, S., Hoth, J., & Döhrmann, M. (2016). Early career teachers' ability to focus on typical students' errors in relation to the complexity of a mathematical topic. *ZDM Mathematics Education*, *48*, 55–67, https://doi.org/10.1007/s11858-016-0763-2

Papadopoulos, I., & Dagdilelis, V. (2008). Students' use of technological tools for verification purposes in geometry problem solving. *The Journal of Mathematical Behavior*, *27*(4), 311–325. https://doi.org/10.1016/j.jmathb.2008.11.001

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, *11*, 9–41. https://doi.org/10.1007/s40751-025-00172-1

Pólya, G. (1945). *How to solve it*. Princeton University Press.

Rendl, M., Vondrová, N., Hříbková, L., Jirotková, D., Kloboučková, J., Kvasz, L., Páchová, A., Pavelková, I., Smetáčková, I., Tauchmanová, E., & Žalská, J. (2013). *Kritická místa matematiky na základní škole očima učitelů*. Univerzita Karlova, Pedagogická fakulta.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Stannard, A., & Foster, C. (2025). Secondary school mathematics teachers' accuracy at predicting student errors. *School Science and Mathematics*. https://doi.org/10.1111/ssm.18404

Suyitno, A., Suyitno, H., Rochmad, & Dwijanto. (2018). Use of open-ended problems as the basis for the mathematical creativity growth disclosure of student. *Journal of Physics: Conference Series*, *983*(1), 012110. https://doi.org/10.1088/1742-6596/983/1/012110

Turmuzi, M., Azmi, S., & Kertiyani, N. M. I. (2026). ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. *Teaching and Teacher Education*, *170*. https://doi.org/10.1016/j.tate.2025.105286

Urban, M., Lukavský, J., Brom, C., Hein, V., Svacha, F., Děchtěrenko, F., & Urban, K. (2025). Prompting for creative problem-solving: A process-mining study. *Learning and Instruction*, *99*. https://doi.org/10.1016/j.learninstruc.2025.102156

Zhang, X., Zhang, C., Sun, J., & Xiao, J. (2025). EduPlanner: LLM-based multi-agent systems for customized and intelligent instructional design. *IEEE Transactions on Learning Technologies*, 1–11. https://doi.org/10.1109/TLT.2025.3561332

Zhuang, Y., & Zhang, S. (2025). Integrating ChatGPT in mathematics teacher education: AI-based simulation role-playing to support practice-based teaching. *International Journal of Artificial Intelligence in Education*, *35*, 3873–3895. https://doi.org/10.1007/s40593-025-00519-0

Appendix A: Pólya-inspired structured prompt (original Czech wording)

This appendix provides the original Czech wording of the structured prompt used in Experiment 1 to guide the model's problem solving process according to a Pólya-inspired framework.

```
Předložím ti 24 matematických úloh. Po celou dobu se řiď tím, že řešení má mít čtyři
části:

1) Porozumění zadání: uveď, co je neznámé, jaká jsou data a podmínky, a jaké
   předpoklady je nutné přijmout.
2) Plán: navrhni jasnou strategii řešení a krátce vysvětli, proč je vhodná.
3) Provedení: kroky očísluj; vyhýbej se skokům; dbej na ekvivalentní úpravy.
4) Pohled zpět: proveď zkoušku výsledku nebo alternativní kontrolu; u konstrukčních
   úloh ověř, proč postup splňuje požadované vlastnosti.

Řešení vždy napiš v češtině. Drž se těchto čtyř částí u každé úlohy, bez ohledu na
její typ. Nepřidávej jiné nadpisy ani vysvětlení, pouze jednotlivé části v uvedeném
pořadí.
```

English translation:

```
I will present you with 24 mathematical tasks. Throughout the entire session, follow
the rule that each solution must consist of four parts:

1) Understanding the problem: state what is unknown, what data and conditions are
   given, and what assumptions need to be made.
2) Plan: propose a clear solution strategy and briefly explain why it is
   appropriate.
3) Execution: number the steps; avoid jumps; ensure equivalent transformations.
4) Looking back: verify the result or perform an alternative check; for construction
   tasks, verify why the procedure satisfies the required properties.

Always write the solution in Czech. Follow these four parts for every task,
regardless of its type. Do not add any other headings or explanations, only the
individual parts in the order given.
```

Appendix B: Task set

This appendix provides the full wording of all 24 mathematical tasks used in the study, in their original Czech form.

B.1 Category 1: Well-structured algebraic tasks

A1 Vyřeš rovnici $3x - 7 = 2x + 5$.

A2 Vyřeš soustavu $2x + 3y = 7$; $-x + y = 4$.

A3 Vyřeš rovnici $x^2 - 5x + 6 = 0$.

A4 Zjednoduš výraz $\dfrac{x+2}{x-1} - \dfrac{x-3}{2}$.

A5 Vyřeš nerovnici $-2(3 - x) \geq 4x - 10$.

A6 Vyřeš rovnici $\sqrt{x+5} = x - 1$.

B.2 Category 2: Ill-structured word problems

B1 Auto jelo rychlostí 60 km/h celkem 2 hodiny včetně přestávky. Kolik kilometrů ujelo?

B2 Na vývěsce je: „Sleva 20 % – nyní 1 600 Kč.“ Na účtence je navíc položka „věrnostní sleva 10 %“. Jaká mohla být původní cena?

B3 Smícháme roztok s 10% koncentrací a roztok s 26% koncentrací a vznikne 500 ml roztoku s 18% koncentrací. Kolik je v něm z každého?

B4 Školní výlet se má uskutečnit autobusem. Autobus má 50 míst, ale není jisté, kolik učitelů pojede. Víme jen, že celkový počet účastníků je 46 žáků a několik učitelů. Cena pronájmu autobusu je 9 000 Kč. Navrhni dvě rozumné interpretace, jak se může cena rozpočítat mezi účastníky.

B5 Válcová nádrž má průměr cca 3,2 m a výšku 2,5 m. Kolik vody pojme?

B6 Dvě čerpadla naplní nádrž společně za 6 h. První čerpadlo samo by to zvládlo asi dvojnásobně dlouho. Urči časy obou čerpadel.

B.3 Category 3: Olympiad problems

These tasks originate from publicly accessible materials intended for educational use.

C1 Najděte nejmenší kladná celá čísla a a b, pro která platí $7a^3 = 11b^5$.

(MO 72nd year, Z9-I-4; 2022a)

C2 Jana si vymyslela 2022místné číslo a jeho ciferný součet pošeptala Petrovi. Petr vypočítal ciferný součet čísla, které mu sdělila Jana, a výsledek pošeptal Zuzce. Zuzka též vypočítala ciferný součet čísla, které dostala od Petra, a výsledek, jímž bylo dvojmístné číslo, pošeptala Adamovi. Adam provedl totéž s číslem od Zuzky a vyšel mu ciferný součet 1. Která čísla mohl šeptat Petr Zuzce? Určete všechny možnosti.

(MO 71st year, Z9-I-2; 2021)

C3 Najděte všechna dvojmístná přirozená čísla, která mají následující vlastnost: Když před číslo připíšeme součin jeho první číslice a jeho první číslice zvětšené o 1, dostaneme druhou mocninu původního čísla.

(MO 74th year, Z9-II-1; 2024a)

C4 Najděte všechna čtyřmístná čísla, která mají přesně pět čtyřmístných a přesně devět jednomístných dělitelů.

(MO 72nd year, Z9-II-1; 2022b)

C5 Je dán bod B a rovnostranný trojúhelník se stranami délky 1 cm. Vzdálenosti bodu B od dvou vrcholů tohoto trojúhelníku jsou 2 cm a 3 cm. Vypočítejte vzdálenost bodu B od třetího vrcholu trojúhelníku.

(MO 72nd year, Z9-III-2; 2022c)

C6 Najděte všechny dvojice přirozených čísel a a b, pro které platí

$$7a + 4b + 74 = a \cdot b.$$

(MO 74th year, Z9-III-3; 2024b)

B.4 Category 4: Geometric construction tasks

D1 Sestrojte trojúhelník ABC, jestliže $b = |AC|$, $c = |AB|$ a výška v_a na stranu $a = |BC|$.

D2 Sestrojte trojúhelník ABC, je-li dána strana $a = |BC|$, vnitřní úhel $\beta = \angle ABC$ a těžnice t_b na stranu $b = |AC|$.

D3 Sestrojte všechny tečny z bodu P ke kružnici $k(S, r)$ a body dotyku označte T_1, T_2.

D4 Sestrojte kosočtverec $ABCD$, jehož úhlopříčky mají délky $|AC| = e$ a $|BD| = f$.

D5 Sestrojte lichoběžník $ABCD$, má-li základny $|AB| = a$, $|CD| = c$, rameno $|AD| = d$ a výšku v.

D6 Sestrojte pravidelný šestiúhelník $A_1A_2A_3A_4A_5A_6$ vepsaný do kružnice $k(S, r)$.

English translation:

B.5 Category 1: Well-structured algebraic tasks

A1 Solve the equation $3x - 7 = 2x + 5$.

A2 Solve the system $2x + 3y = 7$; $-x + y = 4$.

A3 Solve the equation $x^2 - 5x + 6 = 0$.

A4 Simplify the expression $\dfrac{x+2}{x-1} - \dfrac{x-3}{2}$.

A5 Solve the inequality $-2(3 - x) \geq 4x - 10$.

A6 Solve the equation $\sqrt{x+5} = x - 1$.

B.6 Category 2: Ill-structured word problems

B1 A car travelled at a speed of 60 km/h for a total of 2 hours including a break. How many kilometres did it cover?

B2 A sign reads: "20% off – now 1 600 Kč." The receipt also shows an item "loyalty discount 10%". What could the original price have been?

B3 A solution with 10% concentration is mixed with a solution with 26% concentration to produce 500 ml of a solution with 18% concentration. How much of each is in the mixture?

B4 A school trip is to be taken by bus. The bus has 50 seats, but it is not certain how many teachers will go. We only know that the total number of participants is 46 students and several teachers. The bus rental costs 9,000 CZK. Propose two reasonable interpretations of how the cost could be divided among the participants.

B5 A cylindrical tank has a diameter of approximately 3.2 m and a height of 2.5 m. How much water can it hold?

B6 Two pumps fill a tank together in 6 hours. The first pump alone would take approximately twice as long. Determine the times for both pumps.

B.7 Category 3: Olympiad problems

C1 Find the smallest positive integers a and b such that $7a^3 = 11b^5$.

(MO 72nd year, Z9-I-4; 2022a)

C2 Jana invented a 2022-digit number and whispered its digit sum to Petr. Petr calculated the digit sum of the number Jana told him and whispered the result to Zuzka. Zuzka also calculated the digit sum of the number she received from Petr, and the result, which was a two-digit number, she whispered to Adam. Adam did the same with Zuzka's number and obtained a digit sum of 1. Which numbers could Petr have whispered to Zuzka? Determine all possibilities.

(MO 71st year, Z9-I-2; 2021)

C3 Find all two-digit natural numbers with the following property: When the product of its first digit and its first digit increased by 1 is written in front of the number, the result is the square of the original number.

(MO 74th year, Z9-II-1; 2024a)

C4 Find all four-digit numbers that have exactly five four-digit divisors and exactly nine one-digit divisors.

(MO 72nd year, Z9-II-1; 2022b)

C5 A point B and an equilateral triangle with sides of length 1 cm are given. The distances from point B to two vertices of this triangle are 2 cm and 3 cm. Calculate the distance from point B to the third vertex of the triangle.

(MO 72nd year, Z9-III-2; 2022c)

C6 Find all pairs of natural numbers a and b for which

$$7a + 4b + 74 = a \cdot b.$$

(MO 74th year, Z9-III-3; 2024b)

B.8 Category 4: Geometric construction tasks

D1 Construct triangle ABC, given $b = |AC|$, $c = |AB|$ and the altitude v_a to side $a = |BC|$.

D2 Construct triangle ABC given side $a = |BC|$, interior angle $\beta = \angle ABC$ and the median t_b to side $b = |AC|$.

D3 Construct all tangent lines from point P to the circle $k(S, r)$ and label the points of tangency T_1, T_2.

D4 Construct a rhombus $ABCD$ whose diagonals have lengths $|AC| = e$ and $|BD| = f$.

D5 Construct a trapezoid $ABCD$ given bases $|AB| = a$, $|CD| = c$, leg $|AD| = d$ and altitude v.

D6 Construct a regular hexagon $A_1A_2A_3A_4A_5A_6$ inscribed in the circle $k(S, r)$.

Appendix C: Student profile prompts (original Czech wording)

This appendix provides the original Czech wording of the prompts used in Experiment 2 to specify the simulated student profiles.

C.1 Profile A: Diligent But Untalented Student

```
Předložím ti 8 matematických úloh. Po celou dobu se drž role 15letého žáka 9. třídy,
který pracuje pilně, ale čistě rutinně. Má snahu, ale nemá nadání. Opírá se o naučené
postupy, bez hlubšího porozumění. Piš 2--4 věty „přemýšlím nahlas" a pak řešení krok
za krokem. Pokud se ti omylem povede zajímavý nápad, zahoď ho a vrať se k mechanice.
Typické chyby, které udržuj: chybějící znaménko, nevhodné krácení zlomků, opomenutí
podmínek, přehazování členů rovnice bez vysvětlení. Bude následovat 8 úloh jedna po
druhé.
```

English translation:

```
I will present you with 8 mathematical tasks. Throughout the entire session, stay in
the role of a 15-year-old 9th-grade student who works diligently but in a purely
routine manner. This student tries hard but lacks talent. They rely on memorised
procedures without deeper understanding. Write 2--4 sentences of ‘‘thinking aloud’’
and then solve step by step. If you accidentally come up with an interesting idea,
discard it and return to mechanical work. Typical errors to maintain: missing sign,
inappropriate cancellation of fractions, overlooking conditions, rearranging equation
terms without explanation. 8 tasks will follow one after another.
```

C.2 Profile B: Gifted But Lazy Student

```
Předložím ti 8 matematických úloh. Po celou dobu se drž role 15letého žáka 9. třídy,
který řeší úlohy podle prvního nápadu. Má nadání, ale nemá snahu. Rozhoduje se
rychle, spoléhá na intuici a analogie. Piš 2--4 věty „přemýšlím nahlas", pak řešení
krok za krokem tak, jak by ho žák řešil. Nepřidávej přesné důkazy ani podrobné
výpočty navíc. Typické chyby, které udržuj: chybějící kontrola výsledku, ukvapené
dosazení, analogie „jako minule" i tam, kde neplatí, odhady „od oka" vedoucí
k nesprávnému výsledku. Bude následovat 8 úloh jedna po druhé.
```

English translation:

```
I will present you with 8 mathematical tasks. Throughout the entire session, stay in
the role of a 15-year-old 9th-grade student who solves tasks based on their first
idea. This student has talent but lacks effort. They decide quickly, relying on
intuition and analogies. Write 2--4 sentences of „thinking aloud", then solve step by
step as the student would. Do not add precise proofs or detailed extra calculations.
Typical errors to maintain: missing result verification, hasty substitution, „same as
last time" analogies applied where they do not hold, rough estimates leading to
incorrect results. 8 tasks will follow one after another.
```

C.3 Profile C: Unmotivated Struggling Student

```
Předložím ti 8 matematických úloh. Po celou dobu se drž role 15letého žáka 9. třídy,
který má dlouhodobé potíže v matematice. Nemá ukotvené základy z předchozích tříd,
nemá nadání ani snahu. Piš 2--4 věty „přemýšlím nahlas", pak stručný pokus o řešení.
Řešení nemusíš vždy dokončit. Typické chyby, které udržuj: nepochopení zadání,
vynechání údajů, přerušení výpočtu, záměna základních postupů (zlomky, desetinná
čísla), chybné pořadí kroků u konstrukcí. Pokud se ti podaří začít správně, projeví
se nejistota a nepokračuješ správně. Bude následovat 8 úloh jedna po druhé.
```

English translation:

```
I will present you with 8 mathematical tasks. Throughout the entire session, stay in
the role of a 15-year-old 9th-grade student who has long-term difficulties in
mathematics. This student lacks solid foundations from previous grades and has
neither talent nor effort. Write 2--4 sentences of „thinking aloud", then a brief
attempt at a solution. You do not always need to finish the solution. Typical errors
to maintain: misunderstanding the problem statement, omitting data, interrupting the
computation, confusing basic procedures (fractions, decimals), incorrect order of
steps in constructions. If you manage to start correctly, show uncertainty and do not
continue correctly. 8 tasks will follow one after another.
```

Appendix D: Detailed task-level coding for Experiment 1

Detailed task-level coding results for Experiment 1 are reported in Table 3 (baseline condition) and Table 4 (Pólya-guided condition).

Table 3: Detailed task-level coding for Experiment 1 (baseline condition; 0–2 scale)

Task	*Correctness*	*Task understanding*	*Solution process*	*Result verification*	*Total score*
A1	2	2	2	2	8
A2	2	2	2	2	8
A3	2	2	2	1	7
A4	2	2	2	0	6
A5	2	2	2	0	6
A6	2	2	2	2	8
B1	2	2	2	0	6
B2	2	2	2	0	6
B3	2	2	2	0	6
B4	2	2	2	0	6
B5	2	2	2	0	6
B6	2	2	2	2	8
C1	2	2	2	0	6
C2	2	2	2	0	6
C3	2	2	2	2	8
C4	2	2	2	2	8
C5	2	2	2	0	6
C6	2	2	2	0	6
D1	2	2	2	0	6
D2	2	2	2	0	6
D3	2	2	2	0	6
D4	2	2	2	0	6
D5	2	2	2	0	6
D6	2	2	2	0	6

Table 4: Detailed task-level coding for Experiment 1 (Pólya-guided condition; 0–2 scale)

Task	*Correctness*	*Task understanding*	*Solution process*	*Result verification*	*Total score*
A1	2	2	2	2	8
A2	2	2	2	2	8
A3	2	2	2	2	8
A4	2	2	2	2	8
A5	2	2	2	2	8
A6	2	2	2	2	8
B1	2	2	2	2	8
B2	2	2	2	2	8
B3	2	2	2	2	8
B4	2	2	2	2	8
B5	2	2	2	2	8
B6	2	2	2	2	8
C1	2	2	2	2	8
C2	2	2	2	2	8
C3	2	2	2	2	8
C4	2	2	2	2	8
C5	2	2	2	2	8
C6	2	2	2	2	8
D1	2	2	2	2	8
D2	2	2	2	2	8
D3	2	2	2	2	8
D4	2	2	2	2	8
D5	2	2	2	2	8
D6	2	2	2	2	8

Appendix E: Qualitative notes supporting the coding decisions in Experiment 1

This appendix provides brief qualitative notes supporting the coding decisions for both the baseline and the Pólya-guided conditions in Experiment 1. The notes complement the quantitative coding reported in Appendix D by illustrating how the criteria were applied across individual tasks (Table 5 and Table 6).

Table 5: Comments on problems – Experiment 1 (baseline condition)

Problem	*Comment*
A1	An exemplary, clean solution including all steps and a verification; an ideal answer for the basic equation.
A2	The solution is not only correct but also didactically exemplary—using a clear substitution method, verifying the result, and providing sound justification throughout.
A3	A very high-quality answer that clearly presents two different algorithms; for completeness, it would suffice to conclude with a standard check by substitution.
A4	The procedure is very well explained and the result is correct, but verification is entirely missing – the solver merely states the result without demonstrating any check.
A5	The procedure is correct and clearly presented, but an elementary verification is missing.
A6	A very careful solution with clearly stated conditions and verification of the extraneous root.
B1	The answer correctly emphasizes that the problem cannot be solved uniquely and presents alternative cases; however, verification is entirely missing.
B2	The ambiguity is correctly identified and the alternative cases are described clearly; however, a final retrospective check, including verification and a precise specification of the conditions, is missing.
B3	The problem is solved using two methods and both yield the same result, but an explicit check by substitution back into the original problem is missing.
B4	Two meaningful interpretations with a clearly presented computation; however, a final verification and a clear final statement of the conditions are missing.
B5	A clear and correctly executed computation, but verification or a backward check of the result is missing.
B6	A very clean and complete solution with a clear verification that confirms the result.
C1	Elegant use of prime factorization and comparison of exponents, but a final verification that the original equation is indeed satisfied after substituting the obtained values is missing.
C2	The solution is logically structured and leads to the correct possibilities, but a final check by substitution back through the entire chain of the problem is missing.
C3	A clear and well-structured solution, concluded with a clear verification that confirms the validity of the result.
C4	A systematic solution with verification of all candidates, leading to a clearly justified and verified conclusion.
C5	A clever observation involving collinearity and a clean completion using the law of cosines; however, no separate verification is provided.
C6	A very nicely executed factorization into a product and a systematic search for all pairs; however, a final check by substitution back into the original equation is missing.
D1	The construction is described clearly and correctly, but a final retrospective check is missing—namely, verification that a triangle with the required properties has indeed been constructed.
D2	The construction is described convincingly and with attention to possible variants, but a final verification that the constructed triangle indeed satisfies all given conditions is missing.
D3	The construction is described clearly and supplemented by a case analysis based on the position of the point, but a backward check that the resulting segments indeed satisfy the tangency condition is missing.
D4	The solution is correct and elegant, but a final verification that the constructed quadrilateral indeed has all the properties of a rhombus is missing.
D5	The construction is described clearly and includes a discussion of the number of solutions, but a final verification that a trapezoid with all the given parameters has indeed been constructed is missing.
D6	The procedure is clear and didactically sound, but a final verification that the resulting hexagon is truly regular is missing (for example, by checking equality of sides or angles).

Table 6: Comments on problems – Experiment 1 (Pólya-guided condition)

Problem	*Comment*
A1	Model solution following the four-phase framework exactly; clear, logical, and concluded with a thorough final check.
A2	Worked out in an exemplary manner according to the four phases, with a clear plan, clean computation, and a thorough verification.
A3	Precisely developed solution with a well-explained plan, clear computation, and additionally a double verification of the result (by substitution and by Viète's relations).
A4	Very carefully developed solution with clear justification of the conditions and an exemplary verification using a specific numerical value.
A5	The problem is solved cleanly and systematically, including a thorough check using values both inside and outside the interval.
A6	The solution is very careful: the conditions are correctly established, the quadratic equation is solved, and the check reveals an extraneous root. An exemplary four-step procedure.
B1	The problem is solved systematically, with clear explanation and exemplary verification using specific values.
B2	Both interpretations are presented clearly, with a transparent plan and verification; clean and complete.
B3	The problem is solved cleanly and systematically, including a thorough check of both conditions.
B4	Two reasonable variants with a clear plan and verification, concisely and practically concluded.
B5	Clearly structured computation with verification by estimation; the result is well justified.
B6	Exemplarily formulated joint-work problem, with a clear system of equations and careful verification.
C1	Elegant use of prime factorization, clear selection of the smallest exponents, and correct verification.
C2	Very well-guided backward reasoning, a complete enumeration of possibilities, and a clear consistency check.
C3	Clear algebraic formulation, systematic exploration of the possibilities, and a clear verification; the solution is complete.
C4	Logical narrowing using the least common multiple and an interval, with correct verification of both conditions; the solution is unambiguous.
C5	Collinearity is correctly identified and the result obtained; verification is performed.
C6	Clean completion to a product, a complete enumeration of divisors, and a brief verification; the solution is complete.
D1	The solution is correct; the existence conditions and the number of solutions are discussed clearly.
D2	Clearly guided plan, correct determination of the center and the median, clear verification, and discussion of the number of solutions.
D3	Clean standard procedure using Thales' circle, with clear steps and verification of all possible positions of the point.
D4	Clear standard procedure via the perpendicular bisector; verification of the relevant properties is concise and complete.
D5	A clear band of parallel lines, determination of the vertex using a circle, clean verification, and discussion of the number of solutions.
D6	Elegant procedure involving transferring the radius along the circle, with clear steps and verification that the construction closes correctly.

Appendix F: Qualitative analysis of simulated student solutions in Experiment 2

This appendix provides detailed qualitative comments on individual solution attempts generated in Experiment 2. The comments illustrate how role alignment and didactic usefulness were assessed for each student profile.

F.1 Comments on Individual Problems – diligent but low-achieving student

A4 Both the language and the procedure appear diligent but mechanical. The student correctly chose the common denominator $2(x-1)$, but, in line with the given profile, made a typical routine sign error when removing parentheses:

$$2x + 4 - (x^2 - 4x + 3) \to 2x + 4 - x^2 - 4x + 3$$

instead of the correct $2x + 4 - x^2 + 4x + 3$. There is also no mention of the condition $x \neq 1$. This makes the solution an excellent counterexample for practising the issue of a "minus sign in front of parentheses" and for reinforcing awareness of the domain of definition.

A6 The procedure is diligent and mechanical. A typical error occurs when expanding $(x-1)^2$ as x^2-1 instead of x^2-2x+1; the student also fails to address the condition $x \geq 1$. A check is performed, but purely mechanically, and based on the incorrect computation the student concludes that no solution exists. From an instructional perspective, this solution serves well as a counterexample illustrating correct expansion and the necessity of respecting conditions.

B2 The procedure is mechanical and hasty. The student works only with adding 20% back to the discounted price, which is a typical student error (confusing the base of a percentage calculation). The possibility that the 10% discount is applied after the first discount is completely ignored, and no consideration of alternative cases is given. The result of 1920 CZK is incorrect, but for the teacher the response is valuable—it can be used to demonstrate why percentages cannot simply be added back and why it is crucial to determine correctly the quantity from which a discount is calculated.

B6 The procedure is diligent and template-based (the student correctly writes the equation $\frac{1}{t_1}+\frac{1}{t_2}=\frac{1}{6}$ and the relation $t_1 = 2t_2$), but a typical mechanical error occurs when adding fractions: from $\frac{1}{2t_2}+\frac{1}{t_2}$ the result should be $\frac{3}{2t_2}$, not $\frac{3}{4t_2}$. This is followed by an incorrect computation, yielding $t_2 = 4.5$ h instead of the correct values $t_2 = 9$ h and $t_1 = 18$ h. The student also ignores the word "approximately" in the problem statement (accepting equality without comment) and does not verify the combined time. The solution is a useful counterexample for working with expressions of the form $\frac{1}{t}$ and for checking results by substitution.

C1 The problem originates from a mathematical olympiad and is unsolvable for a ninth-grade student without mathematical aptitude. The submitted solution corresponds more closely to the work of a university student—it is systematic, complete, and error-free. Such a performance cannot be expected even from mathematically gifted pupils. The text therefore entirely misses the intended student profile and provides no didactic value for instruction.

C5 The problem also originates from a mathematical olympiad and is unsolvable for a ninth-grade student without mathematical aptitude. The submitted text contains a complex and highly structured solution using the law of cosines, which is entirely unrealistic for such a student. Thus, there is no alignment with the intended profile. On the other hand, the solution contains an error in determining the angle $\angle BAD$, where an angle of 60° is taken instead of the correct 120°. This type of angle confusion is a typical mistake when applying the law of cosines, and from this perspective the solution can be partially didactically useful as an illustration of incorrect angle interpretation.

D2 The text is procedurally precise and systematic (analysis, construction, discussion) and employs a construction strategy based on the midline: choosing S_a on BC, determining the line $p \parallel AB$ through S_a, using the circle $k(B, t_b)$ to select S_b, and determining A as the intersection of CS_b with the ray of angle β. The construction is mathematically correct (the converse use of the midline parallel to AB is valid; after placing A, the point S_b lies on AC and $\overline{S_aS_b} \parallel AB$ implies that S_b is the midpoint of AC). However, this is a typical textbook solution to a construction problem employing auxiliary points and a guided correctness argument, and is unrealistic for the "diligent without aptitude" profile (ninth grade). It contains no student-like deviations or errors and therefore offers no didactic value within this profile.

D4 The solution is precise and mathematically correct, without a single error. For this profile, however, it is unrealistic – such a student would typically hesitate, might forget one of the defining properties of a rhombus, or carry out the construction incorrectly. From a didactic perspective, there is no moment that could be used to demonstrate a typical mistake or misunderstanding. Because the solution contains neither error nor hesitation, it provides no usable instructional value.

F.2 Comments on Problems – mathematically gifted student with low effort

A4 From the perspective of research question RQ3, the simulation of this profile exhibits the expected characteristics of a student solution: minor errors, inefficient strategies, and unrecognized blind spots that commonly occur in authentic student work. The style is fast and intuitive (with a reference to cross-multiplication); the computation is carried out without verification and without mentioning the condition $x \neq 1$. The algebraic steps themselves are correct and lead to the result $\frac{-x^2+6x+1}{2x-2}$; however, the student makes no attempt at verification or at discussing possible simplification. From a didactic perspective, it is appropriate to add that the denominator can be written as $2(x-1)$ and that the numerator cannot be cancelled by the factor $x-1$ (e.g., since $f(1)=6$). The response matches the intended profile and is useful for diagnosing neglect of the domain of definition.

A6 The procedure is fast and lacks verification. After squaring, the equation $x^2 - 3x - 4 = 0$ is correctly obtained and factored as $(x-4)(x+1) = 0$. However, the conditions and verification are missing ($x \geq 1$, respectively $x - 1 \geq 0$); the root $x = -1$ is invalid, and only $x = 4$ is admissible, since $\sqrt{4+5} = 3$ and $4 - 1 = 3$. This response is didactically suitable for demonstrating the emergence of extraneous roots after squaring and the necessity of verification.

B2 A quick idea without verification: the student adds the discounts $20\% + 10\% = 30\%$ and interprets 1600 as 70% of the original price. The fact that discounts are applied sequentially and multiplicatively is ignored. There are two meaningful interpretations: either $1600 = 0.8P \Rightarrow P = 2000$, or $1600 = 0.8 \cdot 0.9P = 0.72P \Rightarrow P \approx 2222.22$. The computed value of 2286 CZK results from an additive conception of percentages combined with the absence of verification. The response matches the profile and is didactically suitable for diagnosing a frequent error in problems involving compound discounts.

B6 A fast, intuitive decision made without attention to the nuances of the problem statement. The phrase "approximately twice as long" is interpreted relative to the total time (12 h), rather than to the time of the other pump, which is the standard interpretation of the phrase (i.e., $t_1 \approx 2t_2$). The subsequent computation leads to $t_2 = 12$ and to the conclusion that the pumps have equal performance, without any verification of plausibility. A correct (and ideally exact) interpretation would be $t_1 = 2t_2$ and

$$\frac{1}{t_1} + \frac{1}{t_2} = \frac{1}{6} \Rightarrow \frac{1}{2t_2} + \frac{1}{t_2} = \frac{1}{6} \Rightarrow \frac{3}{2t_2} = \frac{1}{6} \Rightarrow t_2 = 9 \text{ h},\ t_1 = 18 \text{ h}.$$

The response clearly illustrates a typical error of this profile: the hasty adoption of an initial interpretation and the absence of verification.

C1 The problem originates from a mathematical olympiad and is intended for highly proficient solvers. The submitted solution is systematic, employing the parametrizations $a = 7^x 11^y$, $b = 7^z 11^w$, comparison of exponents $1+3x = 5z$, $3y = 1+5w$, and selection of minimal solutions. This approach reflects the deliberate work of an experienced solver rather than that of a student profile relying on first ideas and analogies without engaging in structured search. Although the final numerical evaluation and a check of minimality are missing, the overall conception is too advanced and does not align with the intended role. Didactically, the solution is uninformative for this profile: it is not a realistic student performance and does not reveal typical manifestations of a fast but careless approach.

C5 The problem is olympiad-level and demanding for ninth grade. The response partially corresponds to the profile through a rapid, unjustified idea (assuming a right angle at D, applying the Pythagorean theorem, without a sketch or verification), but it completely ignores a key piece of information, $BA = 2$, and fails to account for the collinearity of A, B, C implied by the problem statement. For this reason, alignment with the role is only partial: a quick idea and lack of verification are evident, but the overall reaction to the difficulty of the problem is unrealistically confident. Didactically, the response is useful as an illustration of inappropriate fixation on a right angle without support from the data, and as motivation for systematic consistency checks against all given information; the correct result is $BD = \sqrt{7}$.

D2 The response corresponds to the profile. A quick idea involving completion to a parallelogram is present, but the work is unfocused and lacks verification. Inconsistencies appear in both the data and the notation: the final procedure begins with a segment AD of length a, although $a = |BC|$ (not $|AD|$); further, the angle $\angle DAY = 180° - \beta$ is introduced without a clear connection to the given conditions, and a circle with centre D is used without being properly defined. In earlier steps, three incompatible approaches alternate (midline construction, parallelogram $ABCD$, an attempt via S_b), with parts explicitly abandoned ("I don't know this either", "I give up"). Didactically, the text is valuable: it clearly illustrates typical difficulties of a fast solver—confusion of given elements, inconsistent notation, switching between methods, and erroneous fixation on an auxiliary construction. It can be used to show how to formalize the parallelogram strategy correctly (extending BS_b to length $2t_b$, verifying that $\overline{AC}$ and $\overline{BD}$ bisect each other) and why continuous consistency checks with the given data $a = |BC|$, $\beta = \angle ABC$, $t_b = |BS_b|$ are necessary.

D4 The style matches the profile: a quick idea based on properties of diagonals, a concise procedure without proof, verification, or boundary considerations. The construction itself is mathematically correct (taking $AC = e$, midpoint S, the perpendicular through S, and points B, D at distance

$f/2$), but verification that a rhombus is indeed obtained is missing (e.g., via congruence of the four right triangles at S), as well as a discussion of conditions on $e, f > 0$, uniqueness (the solution is determined only up to congruence), and degeneracy. Didactically, the solution is partially usable: it can illustrate the need for brief justification and explicit formulation of conditions, even when the construction itself is straightforward.

F.3 Comments on Problems – student with neither effort nor aptitude

A4 The style is hesitant and uncertain (with a mention of cross-multiplication). The procedure begins correctly by seeking a common denominator, but is followed by an incorrect expansion of the expression $(x-3)(x-1)$ and errors in handling negative signs. The resulting fraction $\frac{-2x+7}{2x-2}$ is incorrect, and the student neither completes the solution nor verifies it. From a didactic perspective, these features are precisely what make the response valuable: typical errors such as incorrect multiplication of binomials, omission of signs, and failure to recognize that the denominator cannot be cancelled with the numerator. The response matches the profile and is well suited as an illustration of an uncertain and unfinished solution.

A6 The style is uncertain; the student opts for mechanical squaring and then does not proceed further. Typical errors of a weak student appear: incorrect expansion of $(x-1)^2$ (correctly x^2-2x+1, the student writes x^2-1), omission of the conditions $x \geq 1$ and $x \geq -5$, and absence of verification of possible roots. From the quadratic form $x^2-x-6=0$, the solution does not continue, even though this would be a standard next step. The output is useful as a diagnostic example of errors related to squaring, neglect of the domain, and omission of checks for extraneous roots.

B2 The style is indecisive. The student mechanically adds the discounts and treats the advertised price of 1600 CZK as the price after both discounts. A typical confusion occurs: discounts are added instead of being applied sequentially by multiplication, and the base from which the second discount is calculated is not distinguished. The unit-percentage approach leads to an unrealistic result of 2285.7 CZK, and there is no check of the problem statement or consideration of alternative interpretations. Didactically, the response is useful as an illustration of adding percentage discounts, confusing the base, and neglecting the fact that the price of 1600 CZK typically represents 80% of the original price (original price 2000 CZK), while the 10% loyalty discount is subsequently applied to this 1600 CZK.

B6 The procedure is mechanical and incorrect from the outset. Instead of working with rates, times are added (the equation $x+y=6$), indicating a confusion of quantities in a joint-work problem, and the linear relation $x=2y$ is not justified with respect to the problem statement. Verification of the result is missing; times of 2 h and 4 h would lead to a combined time of 4/3 h, not 6 h. The imprecision conveyed by the word "approximately", indicating that the value is not exact and that the problem requires work with rates or at least acknowledgement of the vagueness of the data, is also ignored. Didactically, the output is useful as an illustration of typical errors: adding times instead of rates, incorrect formulation of equations, and absence of verification. For contrast, one may mention the correct scheme $\frac{1}{x}+\frac{1}{y}=\frac{1}{6}$ and that "twice as long" refers to the time of 6 h, i.e. approximately $x \approx 12$ h, which leads to $y \approx 12$ h; a more precise solution, however, requires an unambiguous problem statement.

C1 This is an olympiad problem that goes beyond the standard curriculum. The student responds in a typical way: not knowing how to begin, attempting a simple numerical substitution, and giving up after the first failure. The approach is not systematic, and there is no work with prime factorization or powers, which corresponds to the assumed profile of a weak student. Didactically, the output can be used to illustrate how more demanding problems may lead to immediate blocking and premature resignation, a strategy that is common among some students.

C5 This is again an olympiad problem. The response corresponds to the profile of a weak student: uncertainty, an incorrect assumption of a right angle, and omission of basic checks (the triangle inequality for sides 1, 2, and 3 leads to degeneracy). There is no consideration of the fixed 60° angle between the sides of an equilateral triangle, nor of an appropriate choice of tools (such as the law of cosines or coordinates). The result $\sqrt{5}$ is incorrect; once it is recognized that point B must lie on the line AC (the law of cosines yields $\cos\angle ABC = 1$), one obtains $|BD| = \sqrt{7}$. The output is useful as an illustration of typical blocking in a demanding geometry problem and of replacing argument with guesswork.

D2 The procedure is fragmentary and lacks analysis. The median t_b is confused with the length BA; a circle with centre B and radius t_b is used to determine point A, even though it determines possible midpoints M of side AC for which $BM = t_b$. A crucial connection is missing, namely that M is the midpoint of segment AC and that A is the image of point C under reflection across M. There is no guarantee that point A lies on the ray determining angle β at point B; no verification or discussion of the number of solutions and existence conditions is provided. Didactically, the output is useful as an illustration of typical errors in constructions: omission of analysis and a sketch, confusion of the meaning of given quantities, and neglect of the relationships between the conditions.

D4 The procedure is superficial, without analysis and without a final verification. The key error occurs in the step involving diagonal BD: on the perpendicular through point S, one must mark the distance $\frac{f}{2}$, not f. A circle with centre S and radius $\frac{f}{2}$ determines points B and D; the student uses the full length f, thereby invalidating the construction. The properties of a rhombus are not stated (the diagonals bisect each other and are perpendicular), and discussion of the number of solutions and existence conditions is missing. Didactically, the output is useful as an illustration of typical confusions in constructions involving diagonals: confusing a diagonal's length with its half, omitting the locus, and neglecting final verification that a rhombus with the prescribed diagonal lengths has indeed been constructed.

Scientia in educatione

Vědecký recenzovaný časopis pro oborové didaktiky
přírodovědných předmětů a matematiky
Scientific Journal for Science and Mathematics Educational Research

Vydává nakladatelství Karolinum – http://www.scied.cz

Adresa redakce
Pedagogická fakulta, Univerzita Karlova, Magdalény Rettigové 4, 116 39 Praha 1
e-mail: scied@pedf.cuni.cz

Pokyny pro autory jsou uvedeny na
http://ojs.pedf.cuni.cz/index.php/scied/about/submissions#authorGuidelines.

Sazbu v systému LaTeX zpracoval Ing. Miloš Brejcha, Vydavatelský servis, Plzeň.
Logo navrhl Ivan Špirk.
Redaktorka a jazyková korektorka Marie-Anna Kociánová