From Real-world Contexts to Physics Models: Chinese Senior High School Students' Difficulties in Solving Contextualized Physics Problems

Hao Tai*
Binhai Dongyuan Senior High School, Yancheng 224500, China
*Corresponding email: 516773540@qq.com

Abstract

Contextualized physics problems require more than the application of a familiar law: Students must reconstruct a physical system from ordinary language, distributed information, technical devices, and real-world constraints. This qualitative-dominant embedded multiple-case study examined where that reconstruction breaks down among Chinese senior high school students and how classroom and assessment conditions shape the process. The evidence base comprised 8 expert reviews, 12 student cognitive interviews, 24 task-linked student case records selected by school, grade, and performance stratum from a 420-case sampling frame, 15 physics-teacher interviews, and 4 interviews with research or curriculum leaders. Analysis used a recursively connected six-phase framework: context decoding, information structuring, model and idealization, representational translation, solution and monitoring, and interpretation and validation. Four conditions emerged across sources: examination tempo and process compression, contextual-experience fairness, scaffold dependence and fading, and assessment-evidence misalignment. Among the 24 purposefully selected student cases, representational translation was most frequently assigned as the dominant difficulty (10 cases) and remained visible in every performance stratum. Lower-performing cases showed broader cascades across early decoding, model selection, representation, and monitoring; middle-performing cases often knew the relevant law but lost hidden conditions or multi-stage connections. Higher-performing cases generally repaired early breakdowns but still compressed explanation, model boundaries, and transfer. Cross-case analysis produced a recursive mechanism in which early structuring disruption encourages formula-first shortcuts, weakens model-representation coordination, narrows monitoring to arithmetic, and produces premature numerical closure. Examination pace and answer-centered scoring amplify this cycle, whereas phase-specific prompts, explicit representation mapping, model comparison, reasonableness checks, and deliberate scaffold fading provide interruption points. The study contributes a process-based account of contextualized physics difficulty and argues that assessment should capture the evidence chain from situation to model and interpretation rather than reward only the final answer.

Keywords

Contextualized physics problems, Scientific modeling, Physics problem solving, Multiple representations, Senior high school, Qualitative research

Introduction

Contemporary physics curricula increasingly ask students to use disciplinary knowledge in settings that resemble transport, engineering, household technology, sport, energy use, and public decision making. Such tasks are often described as contextualized, context-rich, authentic, or real-world problems. Their educational promise is substantial: A context can give a quantity a purpose, make assumptions consequential, and require students to interpret an answer rather than merely produce it. At the same time, recent evidence cautions against treating contextualization as automatically motivating. In a large-scale study of secondary students and teachers, Wei et al. found a marked student-teacher perception gap: Teachers generally valued contextualized physics problems, whereas older students often preferred decontextualized exercises and did not consistently experience the context as motivational [1]. This tension suggests that the important question is not whether a problem has a context, but what cognitive and epistemic work the context requires and whether students have learned to perform that work.

Research on authentic science, technology, engineering and mathematics (STEM) education distinguishes superficial realism from epistemically productive authenticity. Authentic tasks do not become educationally meaningful simply by mentioning a familiar device or lengthening the problem stem. They should engage learners in constructing, using, testing, and communicating models in relation to consequential constraints [2]. A related analysis of authenticity in science education emphasizes that authenticity is multidimensional and dependent on how learners participate in scientific practices, not merely on whether a scenario resembles everyday life [3]. For physics problem solving, this distinction is decisive. Lengthy problem narratives describing vehicle braking often reduce to simple formula substitution. Concise scenarios, however, foster genuine model-based thinking. To achieve this outcome, students complete four distinct tasks. They identify the relevant physical system, provide reasoning for overlooking minor factors, integrate multiple mathematical representations, and assess if their results align with physical principles and the practical context of the problem.

The transition from a situation to a model is therefore best understood as sensemaking rather than translation by keyword. Physics sensemaking involves coordinating prior knowledge, representations, mathematical relationships, and checks on coherence. Modeling and sensemaking are analytically distinct but tightly intertwined; barriers in model construction can interrupt sustained sensemaking, while external representations can expose and resolve inconsistencies [4]. Teaching approaches also matter. Epistemological messages that frame models as revisable explanatory tools support more meaningful engagement than instruction that presents modeling as a fixed sequence to be imitated [5]. These perspectives challenge a common classroom pattern in which students search for numerical data, recall a familiar formula, and treat successful substitution as evidence that the physical situation has been understood.

Assessment research shows that modeling competence is multidimensional and task dependent. In Newtonian mechanics, high-school modeling performance has been characterized through conceptualization, deployment of scientific knowledge, and representation, with substantial variation across tasks [6]. Work with preservice science teachers similarly separates meta-modeling knowledge, modeling practice, and modeling product, warning that a correct final model may conceal an impoverished process and that an elaborate process may not yield a correct product [7]. These distinctions are especially relevant to contextualized physics problems because students can obtain partial or even correct numerical answers through fragile routines. A score alone may not reveal whether the student identified the system, constructed appropriate idealizations, mapped diagrams to equations, or understood the model’s domain of validity.

A second body of work concerns the coordination of mathematics and physical meaning. The sci-math sensemaking framework distinguishes forms of reasoning that connect equations with scientific mechanisms rather than treating equations as detached calculation templates [8]. A later cognitive framework describes increasingly sophisticated qualitative, quantitative, and conceptual levels of blended mathematical-scientific sensemaking [9]. In contextualized problems, these forms of sensemaking are required before, during, and after calculation: Variables must be derived from a situation, operations must be connected to physical relationships, and the numerical result must be translated back into the original decision or phenomenon.

Multiple representations are a recurrent bottleneck. Eye-movement research shows that prior knowledge influences how students attend and coordinate text, diagrams, and argumentation [10]. Gaze-transition studies distinguish more and less productive pathways through sequential and simultaneous synthesis problems [11]. The format of representations shapes learners’ conceptual understanding and visual processing capabilities [12]. Meanwhile, a systematic review focusing on eye tracking within physics education draws a clear conclusion. Visual attention data carries meaningful insights only when analyzed alongside task requirements and subject-specific logical reasoning [13]. Collectively, this literature supports a view of diagrams, graphs, free-body diagrams, and equations as components of a coordinated representational system. Merely drawing a picture is not enough; every salient object, direction, stage, variable, and equation term must be mapped consistently.

Recent research in Chinese secondary physics also demonstrates that mathematical resources strongly influence problem-solving performance. A municipal assessment of Grade 12 students identified algebraic and geometric skills as major components of performance [14]. Yet mathematical proficiency is not equivalent to physical modeling. Analyses of students' physics work show that productive mathematics includes unit reasoning, interpretation of variable relationships, and movement between conceptual and calculational orientations [15]. A student can manipulate an equation accurately while using an inappropriate system boundary, overlooking a stage transition, or failing to judge whether the result is plausible.

The instructional environment shapes which parts of the process become visible. An observation protocol for modeling-based science teaching emphasizes not only model construction but also evaluation, revision, use, and classroom discourse [16]. Cross-national classroom studies show that model construction stands out among modeling activities like evaluation and revision. This observation reveals that standard teaching only presents a partial picture of the full scope of modeling work [17]. Scaffolded modeling projects show the value of prompts for sensemaking, articulation, process management, and reflection, but also the need to fade support so that learners can regulate the process independently [18]. Moreover, equity-oriented physics education research argues that observed gaps should not be attributed automatically to students; course structures, assessment opportunities, and institutional routines can produce or remove disparities [19]. Recent work on reflective prompts further shows that evaluation strategies during physics problem solving are not automatically activated and require carefully designed opportunities for use and transfer [20].

Against this background, the present study investigates contextualized physics problem solving as a recursive evidence chain from situation to model, representation, solution, and interpretation. This research targets Chinese senior high school students as participants. This student group lives in a unique learning context. They face heavy exam pressure and have rich experience solving quantitative physics problems. Meanwhile, the latest curriculum lays growing stress on real-life scenarios and discipline-specific practical skills. Rather than asking only which topics students find difficult, the analysis asks where the physical structure is lost, how students compensate for that loss, and how classroom and assessment conditions sustain or interrupt the resulting patterns.

The study makes three related contributions. First, it conceptualizes contextualized physics problem solving as an evidence chain in which everyday meaning must be reorganized into disciplinary objects, stages, assumptions, representations, and tests. Second, it examines how breakdowns propagate across that chain rather than treating decoding, modeling, representation, and checking as independent deficits. Third, it places contextual experience, examination tempo, scaffolding, and scoring practices inside the explanatory account. The resulting framework is therefore neither a linear problem-solving recipe nor a list of student weaknesses. It is a conditional model of how task features and instructional arrangements shape the visibility, stability, and repair of physics reasoning.

Three research questions guided the study. Research question 1: What phase-specific difficulties characterize Chinese senior high school students' movement from real-world contexts to physics models? Research question 2: How do the configurations of these difficulties differ across low-, middle-, and high-task-performance cases? Research question 3: How do contextual familiarity, instructional scaffolding, examination tempo, and assessment practices amplify or mitigate the observed difficulties?

Conceptual framework

The analytical framework treats contextualized problem solving as six recursively connected phases rather than a rigid linear algorithm. The phases are useful for locating evidence, but a proficient solver may move backward and forward among them. A student may revise an interpretation after drawing a diagram, change a system boundary after checking an equation, or return to the context after obtaining an implausible result in Table 1. Expert review in the present study explicitly warned against interpreting the framework as an irreversible sequence; accordingly, the explanatory model includes feedback loops and revision paths.

Phase 1, context decoding, concerns recognition of the phenomenon, relevant objects, question target, and physical meaning of ordinary or technical language. Failure at this phase includes misunderstanding a device term, reading an everyday phrase literally when it has a specific physical implication, scanning immediately for numbers, or being unable to restate what happens in one's own words. Context decoding is not identical to general reading ability. A student may understand each sentence linguistically yet fail to construct a coherent physical event.

Phase 2, information structuring, concerns selecting relevant information, inferring necessary but unstated conditions, distinguishing distractors, and organizing temporal or causal stages. In contextualized tasks, information may be distributed across prose, diagrams, tables, or captions. The students must determine not only which data matter, but when and why they matter. Multi-stage problems are especially demanding because the terminal state of one stage becomes the initial condition of another. A missed connection can make every subsequent calculation internally correct but physically irrelevant.

Phase 3, model and idealization, concerns defining the system, identifying interactions, selecting governing principles, introducing simplifying assumptions, and judging their defensibility. The key distinction is between model selection and task-type matching. Task-type matching uses surface cues to retrieve a formula; modeling connects objects, processes, assumptions, and evidence. Idealization is not simply deleting friction or treating an object as a particle. It requires a reason why a factor is negligible for the question and an awareness of conditions under which the simplification would fail.

Phase 4, representational translation, concerns movement among verbal descriptions, sketches, process diagrams, free-body diagrams, graphs, and equations. Translation is successful when representations are mutually constrained: directions in a diagram match sign in equations, variables refer to the correct object and stage, graph features correspond to physical events, and equation terms can be located in the model. This phase was expected to be especially important in mechanics, where the same situation may require spatial, force, temporal, graphical, and algebraic representations.

Phase 5, solution and monitoring, concerns planning a route, connecting stages, executing mathematical operations, and monitoring units, signs, assumptions, and intermediate results. Monitoring is broader than checking arithmetic. It includes asking whether the chosen model still fits the situation, whether a conserved quantity applies across the specified system, whether a direction convention has changed, and whether a connecting variable has been carried correctly from one stage to the next.

Phase 6, interpretation and validation, concerns returning the result to the context, checking magnitude and limiting cases, explaining implications, and stating model boundaries. A numerical answer with a unit is not yet a contextualized solution. For example, a stopping distance should be interpreted relative to available distance; an acceleration should be judged against the described motion; and a safety decision should include the assumptions on which it depends. This final phase also creates feedback: an implausible result should send the solver back to the context, model, representation, or calculation rather than merely prompt another arithmetic pass.

The framework also distinguishes conditions from process phases. Context familiarity can reduce decoding effort, but familiarity is not evidence of physics ability. Cognitive load and anxiety may interrupt otherwise available knowledge. Teaching support can make invisible processes explicit, but a permanent template can become a substitute for independent judgment. Finally, examination and assessment practices influence which evidence students learn to produce. When only the final answer earns credit and classroom time rewards speed, students receive a rational incentive to compress restatement, model justification, representation mapping, and validation.

Table 1. Six-phase analytic framework for contextualized physics problem solving.

Phase	Core analytic question	Indicative difficulty evidence
D1 Context decoding	What phenomenon, object, target, and physical meaning are expressed?	Terminology confusion; formula/number scanning; inaccurate restatement
D2 Information structuring	Which information matters, what is	Relevant/irrelevant confusion; hidden-

Phase	Core analytic question	Indicative difficulty evidence
	implicit, and how are stages connected?	condition omission; broken stage relation
D3 Model and idealization	What is the system, which interactions matter, and why is a simplification valid?	Unstable boundary; unsupported idealization; task-type matching
D4 Representational translation	How do words, diagrams, graphs, variables, and equations constrain one another?	Translation interruption; variable/direction/stage mismatch
D5 Solution and monitoring	What is the plan, how are stages joined, and what is checked during solution?	Direct substitution; connection error; arithmetic-only checking
D6 Interpretation and validation	What does the result mean, is it reasonable, and where does the model apply?	Number without explanation; no magnitude/limit check; unclear boundary

Method

Design and evidence architecture

A qualitative-dominant, embedded multiple-case design was used to explain how difficulties were organized across phases and conditions. The design integrated instrument-refinement evidence, student task profiles, structured task-linked interview records, teacher accounts, and system-level perspectives. Quantitative profile values were used descriptively to select contrasting cases and contextualize interpretations; they were not treated as representative survey analysis. The 420 valid student cases constituted the sampling frame for the 24 formal student cases. No population estimates, psychometric models, or inferential comparisons are reported from the 420-case frame in this article as shown in Figure 1.

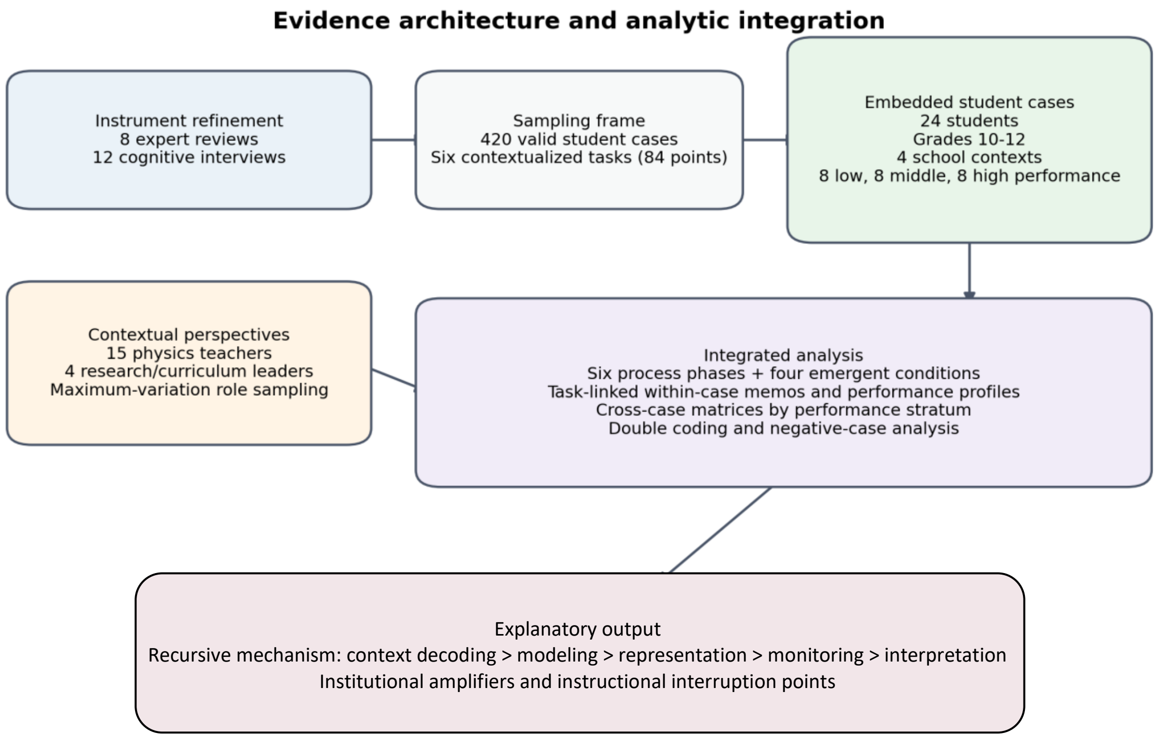

Figure 1. Evidence architecture and analytic integration (The 420 student cases served as the sampling frame; the reported case counts and profiles refer to the 24 purposefully selected formal student cases).

Evidence sources and case selection

The evidence base contained 63 records: 8 expert reviews, 12 student cognitive interviews, 24 formal student case records, 15 physics-teacher interviews, and 4 interviews with research or curriculum leaders. The expert group included university specialists in physics education, curriculum and instruction, qualitative research, and educational measurement, together with municipal or district physics researchers and school curriculum leaders. Reviews addressed construct coverage, wording, scoring feasibility, contextual fairness, and implementation. Cognitive interviews included students from Grades 10-12 with varied self-reported achievement positions; paraphrasing, recall, and response-selection probes were used to identify expressions that could shift the intended construct from physics reasoning toward lexical familiarity.

The student sampling frame was based on six contextualized tasks with a total possible score of 84. The task contexts included shuttle-bus braking, elevator readings, motion on a wet curve, skateboard energy, vehicle collision, and drone release. The 24 formal cases were selected through stratification by school context, grade, and task performance rather than teacher nomination. Four school contexts were represented: an urban demonstration school, an urban regular school, a county regular school, and a township regular school. Grades 10, 11, and 12 were included, with eight cases in each of three performance strata. Each case profile contained the task score and five-point descriptive indicators for perceived difficulty, context familiarity, cognitive load, modeling self-efficacy, and instructional support. These indicators functioned as case descriptors and negative-case prompts, not as validated outcome scales in the present analysis.

The teacher sample comprised 15 senior high school physics teachers from at least six schools and was organized for maximum variation by career stage: five novice teachers, five experienced classroom teachers, and five lead teachers or subject leaders. Reported teaching experience ranged from 1 to 28 years. The four research and curriculum leaders included municipal and district physics researchers and school-level subject or curriculum leaders. Career stage and role were used to contrast perspectives; they were not interpreted as linear measures of professional development.

Table 2. Evidence sources and analytic functions.

Source	n	Composition	Analytic function
Expert reviews	8	Physics education, curriculum, qualitative research, measurement, and school/research leadership	Construct coverage, wording, scoring, fairness, implementation
Cognitive interviews	12	Grades 10-12; varied self-reported achievement positions	Comprehension, retrieval, judgment, and response-process evidence
Formal student cases	24	Four school contexts; Grades 10-12; 8 low, 8 middle, 8 high task performance	Task-linked response records, profiles, and within-case comparison
physics-teacher interviews	15	At least six schools; 5 novice, 5 experienced, 5 lead/subject leaders	Classroom diagnosis, scaffolding, constraints, and fairness
Research/curriculum leaders	4	Municipal/district researchers and school subject/curriculum leaders	Assessment, implementation, resources, and system conditions

Data generation and task-linked records

Each formal student record contained a role and performance profile, responses to eight standardized prompts, initial codes, a field note, and an analytic memo. Students were given a series of prompt questions. They needed to record their initial thoughts when tackling complex context-rich problems. They also had to point out the most challenging part of each task, separate core implicit and unnecessary information, and clarify the defined physical system along with simplifying assumptions. Beyond that, participants matched graphical diagrams with mathematical equations, noted how they tracked their reasoning progress, interpreted calculated outcomes, and listed instructional resources that could improve their problem-solving process. The interview was anchored in the student's task profile and written solution materials. Neutral probes were used to clarify meaning without supplying a correct physical answer.

Teacher interviews proved multiple classroom issues. They explored where students miss physical structure, distinguished modeling barriers from language and experience limitations, and examined the teaching of idealization and representation. The interviews further investigated constraints from teaching pace and class size, individualized scaffolding needs, contextual fairness management, and necessary school and research support. Leader interviews focused on curricular expectations, school-level variation, the evidence captured by assessment, risks of formalistic implementation, contextual fairness, and feasible improvement strategies.

Coding, cross-case analysis, and analytic quality

Analysis began with a codebook aligned to the six phases. Each category contained operational subcodes, inclusion rules, and boundaries that separated, for example, model error from algebraic error and context decoding from final interpretation. An open category was retained for recurrent meanings that challenged or extended the initial framework. Four conditions stabilized during analysis: examination tempo and process compression (NX1), contextual-experience fairness (NX2), scaffold dependence and fading (NX3), and assessment-evidence misalignment (NX4). Coding proceeded at the level of meaning units, followed by within-case memos, cross-case matrices by performance stratum, and comparison across student, teacher, and leader sources.

Analytic quality was strengthened through instrument refinement, task-linked case comparison, double coding, and negative-case analysis. Expert and cognitive interviews were used to revise ambiguous wording and separate contextual familiarity from modeling competence. Student accounts were compared with task scores, weak-task identities, written-solution evidence, and profile indicators. Nine records comprising 136 meaning units were independently double-coded; initial raw agreement was 0.860 and Cohen's kappa was 0.853. This coefficient is interpreted narrowly as evidence of initial-code stability, not as proof that the final thematic explanation is uniquely correct. Negative cases included a low-performing student who reported relatively little difficulty and several middle-performing students who reported high difficulty despite moderate task scores. These cases prevented self-report from being treated as a direct proxy for competence.

Reporting convention and analytic scope

The source material consisted of structured response records rather than discourse-oriented transcripts. The analysis therefore targeted semantic content, task-linked evidence, and code relations rather than conversational form, syntax, or frequency of repeated wording. Identical or near-identical response formulations were not counted as separate corroboration. All records were analyzed in Chinese, and the English manuscript reports analytic paraphrases linked to study codes rather than extended quotations. Figure 1 summarizes the evidence architecture and analytic sequence.

Results

Phase-specific breakdowns from context to model

The central result was not a single misconception but a patterned loss of physical structure across phases. Students often possessed relevant formulas and topic knowledge, yet the route from context to model remained incomplete. Representational translation was assigned as the dominant category in 10 of the 24 formal cases; information structuring was dominant in five, interpretation and validation in four, solution and monitoring in two, model and idealization in two, and context decoding in one. These counts describe the purposefully selected cases rather than prevalence in the 420-case frame. Their analytic value lies in the recurrence of D4 across all three performance strata and in the different ways that earlier and later breakdowns clustered around it.

Context decoding difficulties appeared when technical vocabulary or everyday descriptions were treated as peripheral decoration rather than as information that defined the phenomenon and decision target. Lower-performing records showed number-and-unit scanning before a coherent event had been constructed. Some students could name the textbook topic but could not restate what was physically happening or what the final decision required. Cognitive interviews also showed that apparently ordinary wording could alter the construct. For example, one student interpreted an instruction to exclude information as a request to cross out text rather than to judge whether the information was needed for the current solution. Such evidence motivated wording revisions and demonstrated that an apparent physics difficulty can begin as a mismatch between task language and the intended reasoning operation.

Information structuring was especially visible in multi-stage and multimodal tasks. Students copied all numerical information when they lacked a criterion for relevance, overlooked hidden conditions until an equation failed, or processed prose and diagrams as separate sources rather than complementary constraints. Several cases identified the correct target and governing law but lost the connection variable between stages, such as failing to carry the terminal state of one interval into the initial state of the next. The error was therefore not adequately described as carelessness. It reflected an incomplete temporal representation of the event.

Model and idealization difficulties took two related forms. First, system boundaries were unstable: the focal object, environment, and interacting bodies were mixed or changed implicitly between equations. Second, familiar simplifications were named without a defensible reason. Some middle-performing cases selected among possible laws by counting equations and unknowns. That strategy can support a consistency check after a model has been constructed, but it becomes fragile when it substitutes for identifying the system, process, interaction, and validity conditions.

Representational translation was the most persistent bottleneck. In many low- and middle-performing records, a sketch resembled the setting but did not regulate the equations. Directions in a free-body diagram were not carried into the sign convention, one symbol denoted different stages, or a graph was treated as optional rather than as a representation of change. In several cases the mapping between diagram and equation remained internal and left no inspectable trace in the written solution. This made stage and sign errors difficult to diagnose because the representations did not constrain one another.

High-performing records illustrated a more productive alternative. Diagrams were used to check signs, state variables, stage boundaries, and equation terms; when a stage relation was unclear, the process was redrawn or represented with a velocity-time graph. These practices externalized reasoning and allowed inconsistencies to be detected before final calculation. Nevertheless, D4 remained the dominant category in four high-performing cases. Higher performance therefore reflected greater capacity to detect and repair representational mismatch, not the absence of such mismatch.

Solution and monitoring difficulties were often downstream consequences of earlier modeling problems. Lower-performing cases commonly began calculation without an explicit plan, changed formulas when blocked, and concentrated checks at the end. Multi-stage tasks exposed failures to carry shared quantities or initial-final state relations. When an answer appeared unreasonable, the first response was often recalculation rather than reconsideration of the system boundary, assumption, or representation. Middle-performing cases were more likely to check units and signs, but often only after all calculations. High-performing cases distributed dimensional, sign, connection, and extreme-case checks across stages and traced anomalous results backward through calculation, equation, model, and context.

Interpretation and validation were consistently compressed. Many records ended with a number and unit even when the task required safety, feasibility, or decision judgment. Students sometimes gave a categorical conclusion but did not explain the quantitative margin, test the magnitude, or state the assumptions on which the conclusion depended. Model boundaries were rarely articulated unless explicitly requested. This pattern helps explain why a student can succeed in conventional exercises yet remain vulnerable when a contextualized problem asks for explanation, transfer, or evaluation under changed conditions.

Performance-related configurations

As shown in Table 3, the three performance strata differed less in whether they encountered difficulty than in where the difficulty accumulated and how effectively they repaired it. The low-performance group averaged 14.25 out of 84 on the contextualized tasks. Their mean perceived difficulty was 3.65, mean context familiarity 2.66, cognitive load 3.69, modeling self-efficacy 2.25, and instructional support 3.02. Dominant difficulties were distributed across five of the six phases: three representation cases, two solution-monitoring cases, and one case each in context decoding, information structuring, and model-idealization. This breadth indicates cascading disruption. An early misread or unstructured process often led to an unsupported formula choice, fragile representation, and late arithmetic checking.

The middle-performance group averaged 51.75 out of 84.00, with mean perceived difficulty 3.40, context familiarity 2.88, cognitive load 3.00, modeling self-efficacy 3.22, and instructional support 3.15. Their dominant difficulties concentrated on representation translation (three cases), information structuring (two), interpretation-validation (two), and model-idealization (one). These students commonly knew the relevant physical law and could complete substantial calculation.

Their breakdowns occurred at transitions: Inferring an implicit condition, deciding which stage a datum belonged to, justifying an idealization, or maintaining a consistent mapping across text, diagram, graph, and equation.

The high-performance group averaged 69.88 out of 84.00. Mean perceived difficulty was 1.96, context familiarity 3.88, cognitive load 2.91, modeling self-efficacy 3.92, and instructional support 3.20. Dominant difficulties were limited to representation translation (four cases), information structuring (two), and interpretation-validation (two). These students more often established the model successfully and monitored calculation, but their vulnerabilities shifted downstream. They sometimes relied on efficient internal mappings that were not fully externalized, or they completed a correct solution without articulating the model boundary. Thus, higher performance reduced the breadth of early breakdown but did not eliminate the need for explicit representation and explanation.

The descriptive pattern should not be converted into a causal claim. The 24 cases were selected for contrast, and the profile indicators were self-reported. The negative cases are particularly informative. SIV02 achieved only 23.81% of the task score but reported perceived difficulty of 1.97. In contrast, SIV10 and SIV13, achieved approximately 63.10% and 64.29% while reporting difficulty above 4.20. These cases show that low perceived difficulty may indicate poor monitoring rather than competence, while high perceived difficulty may accompany accurate awareness and productive effort. For assessment and instruction, self-report should therefore be interpreted alongside process evidence and performance.

Table 3. Descriptive profiles and dominant-difficulty distribution of the 24 formal student cases.

Performance stratum	Task score /84	Perceived difficulty	Context familiarity	Cognitive load	Modeling self-efficacy	Instructional support	Dominant difficulty counts
Low (n=8)	14.25	3.65	2.66	3.69	2.25	3.02	D1=1; D2=1; D3=1; D4=3; D5=2; D6=0
Middle (n=8)	51.75	3.40	2.88	3.00	3.22	3.15	D1=0; D2=2; D3=1; D4=3; D5=0; D6=2
High (n=8)	69.88	1.96	3.88	2.91	3.92	3.20	D1=0; D2=2; D3=0; D4=4; D5=0; D6=2

Note: Profile indicators are descriptive five-point case attributes; no inferential comparison is intended.

Figure 2 maps the dominant cognitive difficulty categories of students split into low, middle, and high-performance strata, with eight purposively selected participants per group. The most prominent shared feature across all tiers is the consistent, high frequency of D4 (representation translation), which acts as an enduring cognitive bottleneck for every proficiency level. Notable stratified differences structure the distribution of learning barriers: low-performing students display a broad spectrum of dominant difficulties across nearly all six problem-solving phases, signaling overlapping deficits throughout the entire reasoning workflow. Middle achievers narrow their core struggles to information structuring, representational translation, and result validation, having resolved foundational context decoding and real-time solution monitoring challenges. High-performing learners further limit their persistent difficulties solely to D4 and D6, recording the maximum case count for representation translation at four student samples. This overall pattern reflects a performance-dependent cognitive configuration rather than a sequential developmental trajectory. Greater task proficiency corresponds to fewer, more targeted cognitive breakdowns that lend themselves to targeted intervention; nevertheless. Even top-tier students remain reliant on formal representational externalization and interpretive justification steps, and neither is cognitive demand eliminated by stronger overall performance.

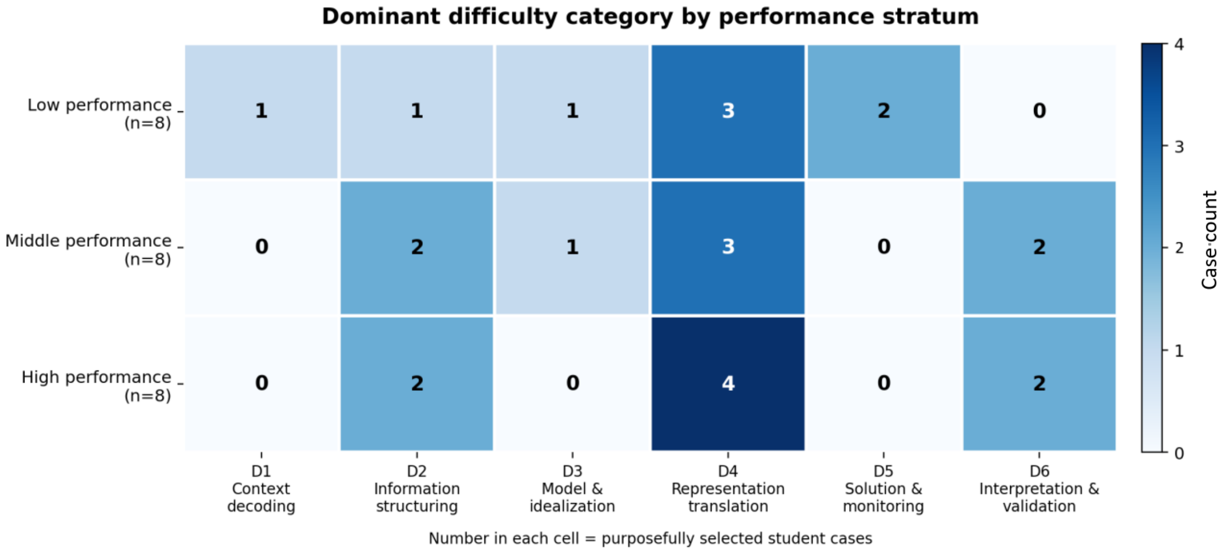

Figure 2. Dominant difficulty category by performance stratum (Cell values are counted among the 24 purposefully selected student cases (n=8 per stratum), not population prevalence estimates).

Case-linked analytic evidence

To make the analytic path visible, Table 4 presents six illustrative cases selected for contrast rather than representativeness. The table links performance profiles, weak tasks, task-linked response evidence, dominant codes, and analytic use. Evidence is paraphrased because the source material was organized as structured response records. The cases show why a total score alone is insufficient: similar scores can arise from different phase breakdowns, and similar self-reports can accompany different levels of monitoring.

The contrast between SIV02 and SIV10 is particularly informative. SIV02 obtained only 20 of 84 points while reporting perceived difficulty of 1.97; the record showed late arithmetic checking and weak transfer of stage variables. SIV10 obtained 54 points but reported difficulty of 4.23 and explicitly recognized uncertainty about hidden conditions and model choice. The former profile is consistent with limited error awareness, whereas the latter suggests that experienced difficulty can accompany active monitoring. SIV23 and SIV24 further show that high performance does not guarantee complete modeling evidence: One case retained a representation mismatch, while the other completed the main reasoning but compressed interpretation and model boundaries.

Table 4. Illustrative case-linked evidence chain.

Case	Profile	Weak task	Task-linked analytic evidence	Dominant category and analytic role
SIV01	Low; 12/84; perceived difficulty 4.20	Vehicle collision	Scanned numbers and units before reconstructing the event; incomplete restatement omitted a stage or decision constraint.	D1 context decoding; illustrates an early cascade
SIV02	Low; 20/84; perceived difficulty 1.97	Shuttle-bus braking	Monitoring concentrated on final arithmetic; stage connections and units were not checked until prompted.	D5 solution/monitoring; negative case for self-report
SIV10	Middle; 54/84; perceived difficulty 4.23	Skateboard energy	Identified the target but treated an implicit condition as given; model choice and idealization remained only partly justified.	D2 information structuring; high awareness with moderate performance
SIV13	Middle; 53/84; perceived difficulty 4.20	Vehicle collision	Information in prose and figures was not fully integrated; signs, variables, and stages were inconsistently mapped.	D4 representation translation; transition failure
SIV23	High; 80/84; perceived	Wet-road curve	Strong system and stage organization coexisted with a	D4 representation translation; residual

Case	Profile	Weak task	Task-linked analytic evidence	Dominant category and analytic role
	difficulty 1.93		diagram-equation mismatch in the weak task.	vulnerability at high performance
SIV24	High; 71/84; perceived difficulty 2.53	Drone release	Calculation and stage checks were generally strong, but the final decision, magnitude, and model boundary were compressed.	D6 interpretation/validation; downstream difficulty

Note: Evidence is reported as analytic paraphrase of the structured Chinese-language case records.

Recursive mechanism of difficulty

Cross-case analysis yielded a recursive mechanism rather than a list of independent errors. The mechanism begins with entry conditions: unfamiliar terminology, limited experience with a device or setting, information density, divided attention across representations, and time pressure. These conditions do not directly determine failure. They increase the probability that students will skip a qualitative reconstruction of the event and move prematurely toward numerical cues.

The first transition is from early disruption to a compensatory formula-first shortcut. When students are uncertain about the target, hidden conditions, or stage relations, they scan numbers, units, and familiar nouns for a recognizable equation. This behavior is understandable in a high-speed examination culture because it sometimes produces partial credit. However, it converts contextualized problem solving into task-type matching. The selected equation is then weakly anchored to a system and process.

The second transition is a model-representation bottleneck. Weak system boundaries and unjustified idealizations make it difficult to decide what a diagram should contain. Conversely, a diagram that is not mapped to variables and equations cannot expose a poor model. The two phases reinforce each other. A student may continue calculating with a mislabeled variable because the drawing is decorative or retain an inappropriate sign convention because the stage change was never represented.

The third transition is calculation-centered monitoring. Once a formula has been selected, attention narrows to substitution and algebra. Checks focus on arithmetic rather than model fit. If the result is implausible, the calculation may be repeated under the same assumptions. The records suggest a reinforcing loop: Unsuccessful reworking can increase experienced load and make a familiar formula-first routine more attractive, particularly when the task is timed and only the answer is visibly rewarded.

The final transition is premature closure. Students report a number and unit but do not return to the original question, compare the result with a realistic range, test a limiting case, or identify the model boundary. Because the solution is not interpreted, the process loses its strongest opportunity for revision. A proficient cycle would use a surprising result to revisit context and model. The observed weak cycle treats completion of calculation as the endpoint.

The mechanism also explains why context familiarity has a complex role. Familiar settings can accelerate decoding and reduce extraneous load, but they can also invite intuitive assumptions that conflict with the intended model. Unfamiliar settings can be productive if the problem provides the necessary background and makes the physical structure explicit. Contextual fairness therefore concerns construct-irrelevant burden: Students should not need privileged exposure to urban transport equipment, specialized technology, or particular leisure activities to access physics. Providing concise background information does not reduce rigor. It protects the intended construct.

The mechanism is recursive and conditional. High-performing students often bypass or repair the early shortcut, but they may still close the process prematurely at interpretation. Process-oriented scaffolds can interrupt the cycle at multiple points: Restatement prompts at D1, information-purpose tables at D2, system and assumption questions at D3, mapping requirements at D4, staged checks at D5, and reasonableness or boundary prompts at D6. Yet scaffolds must be faded. Students repeatedly reported that a template could enable successful imitation while leaving transfer unstable when the template was removed. Figure 3 summarizes the mechanism, its feedback loops, institutional amplifiers, and instructional interruption points.

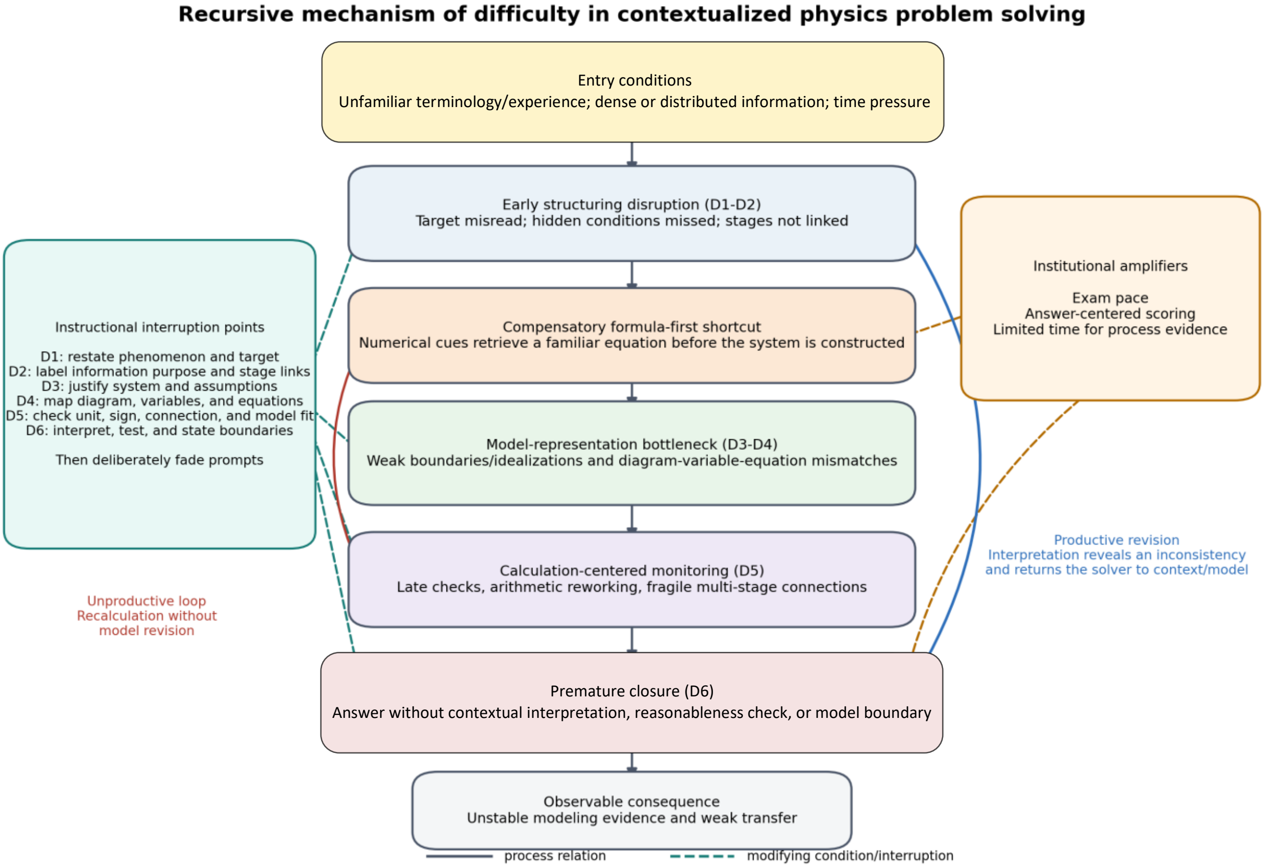

Figure 3. Recursive mechanism of difficulty in contextualized physics problem solving (Solid arrows show the proposed process relation; dashed and curved arrows show modifying conditions and recursive revision or reworking).

Instructional, assessment, and contextual conditions

Teacher interviews converged with the student evidence. Across career stages, teachers described students as moving toward numerical cues and formula retrieval before establishing the object, process, interactions, and stages. They emphasized that middle-performing students often knew the relevant law but lost a connection condition in a multi-stage problem. This distinction is pedagogically important: reteaching the law does not address failure to organize stages or coordinate representations.

Teachers also offered a practical diagnostic for separating language or experience barriers from model barriers. They asked students to restate the situation without formulas and, when necessary, replaced an unfamiliar term with a familiar one. If the student could then describe the event but still could not identify the system and interactions, the difficulty was more plausibly located in modeling. Such diagnosis avoids interpreting every hesitation as low physics ability and aligns feedback with the actual phase of breakdown.

Instructional practice varied in how explicitly it represented modeling. Novice and experienced teachers often report drawing diagrams and writing equations while giving inconsistent explanations for why a factor could be ignored. Lead teachers more often described a visible board sequence object, process, interaction, assumption, representation, test and compared alternative models, such as retaining versus neglecting resistance. The evidence does not establish a causal career-stage progression, but it illustrates practices that make model decisions inspectable.

Across career stages, teachers identified severe implementation constraints. A lesson must cover content and maintain practice volume; classes commonly contain more than forty students; and monthly or entrance-oriented examinations remain dominated by answer scores. Under these conditions, listening to individual explanations or jointly evaluating model boundaries appears costly. The result is process compression: teachers demonstrate a complete solution while students reproduce the visible equations, and the less visible reasoning is omitted.

Teachers differentiated support by student need. Lower-performing students benefited from segmented prompts and partial representations; middle-performing students needed questions about stage connections and hidden conditions. Higher-performing students needed counterexamples, model boundaries, and transfer tasks. The common principle was phase-specific feedback. Telling a student that the answer is wrong does not distinguish context, model, representation, connection, or calculation. Feedback that names the broken link can guide revision and produce data for teaching.

Research and curriculum leaders extended the explanation to school structures. They argued that contextualized tasks should assess an evidence chain, not merely final computation. School differences arose in access to vetted task resources, shared planning time, teachers' assessment literacy, and students' exposure to technological contexts. Leaders warned against two formalistic interpretations: equating contextualization with a longer stem and converting the six phases into a fixed slogan. Both can preserve formula-first behavior while giving it a new surface.

Assessment-evidence misalignment was therefore a central institutional condition. Routine test analysis was commonly organized by knowledge point and item score, while evidence about where the process broke was rarely coded. Leaders recommended stage profiles, representative student work, reusable tasks and rubrics, and small-scale joint scoring before wider use. They also cautioned against ranking schools from such data, because contextual opportunity and resource differences can be mistaken for student ability.

The combined empirical evidence strongly suggests that procedural fairness and analytical rigor are fully compatible. A fair contextualized problem provides the rich background needed to understand the real-world setting, avoids treating a particular life experience as an unstated hidden prerequisite, and still preserves authentic physical constraints, inherent uncertainty, and core decision requirements. Equivalent test versions can use different culturally familiar settings while retaining the identical underlying physical structure. Such thoughtful design allows teachers to clearly distinguish the intended modeling demand from accidental cultural or lexical burden.

Table 5. Cross-source mechanism themes and actionable leverage points.

Theme	Convergent evidence	Instructional/assessment leverage
Formula-first shortcut	Student records prioritize numbers/units; teachers report law retrieval before system construction	Require qualitative restatement and system identification before equations
Hidden stage connections	Implicit conditions or terminal-to-initial state transfers are omitted	Use timelines/process diagrams and label connection variables explicitly
Representation mismatch	Drawings do not constrain signs, variables, stages, or equations	Score relational consistency among diagram, graph, variables, and equations
Arithmetic-only monitoring	Unexpected results trigger recalculation rather than model revision	Embed stage checks for units, signs, assumptions, magnitude, and model fit
Premature closure	Numerical answers are not interpreted and model boundaries remain unstated	Require contextual decisions, reasonableness checks, and changed-condition questions
Process compression	Exam pace and answer-centered scoring suppress visible modeling evidence	Rotate process-focused rubric components and jointly score selected student work
Contextual-experience burden	Specialized terminology or life experience adds construct-irrelevant load	Provide necessary background and test equivalent physical structures across contexts
Scaffold dependence	Templates support immediate completion but transfer weakens after removal	Fade prompts deliberately and vary when and where the tool is needed

Discussion

The findings support a process account of contextualized physics difficulty. Students did not struggle simply because context was unfamiliar or because algebra was weak. Difficulty emerged when contextualization reorganized the task: Relevant information was embedded in ordinary language, data were distributed across representations, assumptions required justification, and the final number had to be interpreted. The recursive model explains why adding a real-world narrative to a conventional exercise can increase burden without increasing authenticity. The context can require additional decoding while the scoring system continues to reward formula retrieval.

The theoretical contribution lies in the coupling of phrases and conditions. Existing research has established the importance of modeling, sensemaking, multiple representations, and validation. The present analysis shows how those elements can become a cascade: disrupted information structure encourages a formula-first shortcut; the shortcut produces weak system boundaries; weak boundaries and poorly mapped representations reduce the quality of monitoring; and narrow monitoring enables premature closure. Examination tempo, contextual experience, scaffolding, and scoring practices alter the probability and visibility of each transition. The six phases should therefore be treated as diagnostic locations within a recursive system, not as independent traits or a mandatory six-step algorithm.

Representational translation was the most persistent bottleneck. This result is consistent with research that treats representations as coordinated epistemic tools rather than interchangeable formats [21]. The present study adds a mechanism: representation errors are often produced upstream by uncertain system boundaries and downstream by calculation-centered monitoring. A free-body diagram cannot regulate an equation if the student has not decided which object is the system; an equation cannot expose a stage error if the diagram uses the same symbol across different states. Representation instruction should therefore emphasize mappings and constraints, not the production of isolated diagrams.

The performance-stratified comparison also refines deficit interpretations. Lower-performing cases exhibited cascading difficulty across phases, but the mechanism was not a lack of effort or a single missing law. Their formula-first strategy was partly adaptive to classroom and examination conditions that reward speed and answers. Middle-performing students often possessed the law and basic mathematics yet failed at connections and justification. High-performing students showed stronger monitoring but still compressed interpretation and model boundaries. These patterns suggest that instruction should not reserve modeling for advanced students. Instead, the same evidence chain can be taught with differentiated prompts and progressively reduced support.

The distinction between perceived difficulty and observed performance is theoretically important. Negative cases showed that confidence or low reported difficulty can coexist with weak task performance, while high reported difficulty can accompany productive monitoring and moderate success. Self-efficacy and cognitive load remain relevant, but they should be interpreted as parts of a regulatory system rather than direct proxies for competence. A student who notices conflict may report greater difficulty precisely because the student is monitoring more deeply.

The study also extends work on authenticity. Context should be viewed as an interface between lived experience and disciplinary modeling [22]. A context is educationally productive when it makes model decisions meaningful and supplies sufficient access to the setting. It is unproductive when it functions as decorative prose, introduces irrelevant cultural knowledge, or hides necessary information behind specialized terminology. This account helps explain the student-teacher perception gap reported in recent research: Teachers may value the intended authenticity, while students experience only additional reading load and uncertain scoring.

Modeling competence research emphasizes multidimensionality and task dependence. The present cases show how those dimensions become coupled during problem solving. A weak conceptualization can generate a representation mismatch; a mismatch can produce a stage-connection error; a calculation error can remain undetected because the result is not interpreted. Assessment that reports only a total score obscures these dependencies. A phase profile is not a replacement for a score, but it provides actionable evidence about where instruction and feedback should intervene.

The institutional findings align with research on modeling-based teaching and course-structure explanations of equity. Students learn what counts as physics from the evidence their classrooms request and reward. If model assumptions, representation mappings, and validation are rarely scored, omitting them is a rational response. Conversely, when rubrics allocate credit to those components and teachers jointly examine student evidence, the hidden curriculum changes. The challenge is to avoid turning the evidence chain into a bureaucratic checklist. Recursive revision, alternative models, and context-sensitive judgment must remain visible.

Finally, the study clarifies the role of scaffolding. Process prompts are useful because they externalize decisions that novices cannot yet regulate. A permanent six-box worksheet, however, can become another task-matching routine. Effective scaffolding should be contingent and faded: first, make the decision visible; then, reduce the cues; vary the context; and require students to decide when the tool is useful. This principle is consistent with scaffolded model-based learning and with evidence that reflective prompts must support both immediate evaluation and transfer.

Implications for task design, teaching, and assessment

First, contextualized-task design should begin with the intended evidence chain. Writers should specify what the student must infer about the phenomenon, which information is essential or intentionally redundant, what system and assumptions are defensible, which representations are needed, how stages connect, and what contextual interpretation validates the answer. Every piece of narrative should have a clear function. Specialized terms should be explained unless knowledge of the term is itself part of the construct.

Second, classroom routines should make modeling decisions public. A compact set of guiding questions can be integrated into worked examples, peer discussions and feedback sessions. This set covers core inquiries: What is happening? What is the system? How many stages are there? What is neglected, and for what reason? How do different representations correspond to one another? What makes a result physically unreasonable? These guiding prompts ought not to follow a fixed template. Varied wording stops students from completing them mechanically without reflection.

Third, representation should be assessed relationally. Credit should be awarded not merely for drawing a diagram but for consistency between diagram, variables, directions, stages, and equations. Students can be asked to annotate where each equation term appears in the representation, explain why a graph is useful, or identify a mismatch in a partially completed solution. Such tasks convert representation into a monitoring resource.

Fourth, assessment rubrics should preserve partial evidence. A feasible analytic rubric can include context and target, information and stage structure, system and assumptions, representation mapping, solution and monitoring, and interpretation and validation. Teachers do not need to score every component on every task. Rotating focal components and jointly scoring a small sample can build assessment literacy without creating unmanageable workload. Reflection prompts should be placed where they can trigger evaluation during the solution and then be faded or varied so that the strategy transfers beyond the prompted task.

Fifth, contextual fairness should be reviewed explicitly. Task developers should ask whether success depends on prior exposure to a device, urban infrastructure, sport, or technology that is unrelated to physics. Necessary background can be supplied through a concise description or diagram, and parallel contexts can be tested for equivalent physical structure. Fairness review should record language and experience demands rather than infer ability from demographic location.

Sixth, professional development should use authentic student work. Instead of discussing modeling only at the level of curriculum slogans, teacher teams can code where a student’s evidence chain breaks, compare alternative feedback, and revise tasks. This creates a direct link among assessment, instruction, and school improvement. The mechanism model can function as a diagnostic hypothesis to be tested and refined, not as a fixed taxonomy.

Limitations and future research

First, several limitations bound the claims. The study is qualitative-dominant and purposefully sampled. The 24 formal student cases were selected for variation from a

420-case frame, but the full 420-case dataset was not analyzed in this article. Frequencies and group mean therefore describe the selected cases and should not be generalized to Chinese senior high school students nationally. The four school contexts provide contrast but do not represent the full diversity of regions, curricula, or examination systems.

Second, the student evidence was preserved as structured response records linked to task profiles and written materials rather than as discourse-oriented transcripts. This design supports semantic coding and cross-case comparison but does not permit claims about interactional detail, hesitation, turn-taking, or the exact language form of students' reasoning. Future studies should retain full audio transcripts and pair them with time-synchronized written or digital traces.

Third, interviews and task artifacts reveal reported and observable reasoning but not every real-time cognitive event. Task-linked recall reduces reliance on general self-description but cannot reproduce the original problem-solving episode. Think-aloud protocols, pen or screen traces, classroom observation, and eye tracking could test the timing and direction of the proposed transitions more directly.

Fourth, performance strata were based on one six-task contextualized set. Modeling competence is task dependent, and a student categorized as high performing in these tasks may struggle in another physics domain. Replication should vary mechanics, electricity, energy, waves, and interdisciplinary contexts and should use equivalent task forms to distinguish task-specific from person-level patterns.

Fifth, teacher accounts were not verified through systematic classroom observation. Claims about pace, feedback, class size, and the visibility of modeling practices represent practitioner perspectives. Video-based studies could examine how frequently the six forms of evidence are elicited, how scoring practices shape student work, and whether phase-specific feedback changes classroom interaction.

Finally, the mechanism is explanatory rather than causal. Its arrows represent recurrent relations supported by cross-case convergence, task-linked records, and negative-case analysis. Intervention studies are needed to test whether representation mapping, model comparison, stage-specific checking, and interpretation prompts improve transfer and whether deliberate scaffold fading prevents template dependence.

Conclusion

Contextualized physics problems require students to do more than place values into equations. They must reconstruct a physical system from language, organize information and stages, justify a model, coordinate representations, regulate calculation, and return the result to the situation. The present study shows that difficulty arises when this evidence chain is compressed. Early uncertainty encourages formula-first shortcuts; weak boundaries and representation mappings make multi-stage reasoning fragile; monitoring narrows to arithmetic; and calculation ends before interpretation or validation. Representational translation remained a central bottleneck even in high-performing cases, whereas lower-performing cases showed broader cascades across phases. Examination pace, answer-centered scoring, and unequal contextual experience amplified the pattern. Phase-specific prompts, model comparison, representation mapping, reasonableness checks, fair context design, and deliberate scaffold fading provide practical interruption points. The goal is not a fixed six-step procedure, but a visible and revisable path from context to model.

Funding

This work was not supported by any funds.

Acknowledgements

The author would like to show sincere thanks to those technicians who have contributed to this research.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Wei, Y., Peng, X., Zhong, Y., Pi, F., Zhai, Y., Bao, L. (2025) Can contextualized physics problems enhance student motivation? *Physical Review Physics Education Research*, 21(2), 020117.

[2] Hallström, J., Norström, P., Schönborn, K. J. (2023) Authentic STEM education through modelling: an international Delphi study. *International Journal of STEM Education*, 10(1), 62.

[3] Schriebl, D., Müller, A., Robin, N. (2023) Modelling authenticity in science education:

modelling authenticity in science education. *Science & Education*, 32(4), 1021-1048.

[4] Sirnoorkar, A., Bergeron, P. D., Laverty, J. T. (2023) Sensemaking and scientific modeling: intertwined processes analyzed in the context of physics problem solving. *Physical Review Physics Education Research*, 19(1), 010118.

[5] Ke, L., Schwarz, C. V. (2021) Supporting students' meaningful engagement in scientific modeling through epistemological messages: a case study of contrasting teaching approaches. *Journal of Research in Science Teaching*, 58(3), 335-365.

[6] Zhai, X. (2022) Assessing high-school students' modeling performance on Newtonian mechanics. *Journal of Research in Science Teaching*, 59(8), 1313-1353.

[7] Göhner, M. F., Bielik, T., Krell, M. (2022) Investigating the dimensions of modeling competence among preservice science teachers: meta-modeling knowledge, modeling practice, and modeling product. *Journal of Research in Science Teaching*, 59(8), 1354-1387.

[8] Zhao, F., Schuchardt, A. (2021) Development of the sci-math sensemaking framework: categorizing sensemaking of mathematical equations in science. *International Journal of STEM Education*, 8(1), 10.

[9] Kaldaras, L., Wieman, C. (2023) Cognitive framework for blended mathematical sensemaking in science. *International Journal of STEM Education*, 10(1), 18.

[10] Wang, Z., Cheng, H., Qin, Q. (2025) The research on video analysis of key motion positions based on deep learning technology. *International Journal of e-Collaboration (IJeC)*, 21(1), 1-16.

[11] Ibrahim, B., Ding, L. (2021) Sequential and simultaneous synthesis problem solving: a comparison of students' gaze transitions. *Physical Review Physics Education Research*, 17(1), 010126.

[12] Susac, A., Planinic, M., Bubic, A., Jelicic, K., Palmovic, M. (2023) Effect of representation format on conceptual question performance and eye-tracking measures. *Physical Review Physics Education Research*, 19(2), 020114.

[13] Hahn, L., Klein, P. (2022) Eye tracking in physics education research: a systematic literature review. *Physical Review Physics Education Research*, 18(1), 013102.

[14] Tong, T., Pi, F., Zheng, S., Zhong, Y., Lin, X., Wei, Y. (2025) Exploring the effect of mathematics skills on student performance in physics problem-solving: a structural equation modeling analysis. *Research in Science Education*, 55(3), 489-509.

[15] Bowers, J., Anderson, M., Beckhard, K. (2024) A mathematics educator walks into a physics class: Identifying math skills in students' physics problem-solving practices. *Journal for STEM Education Research*, 7(3), 335-361.

[16] Shi, F., Wang, L., Liu, X., Chiu, M. H. (2021) Development and validation of an observation protocol for measuring science teachers' modeling-based teaching performance. *Journal of Research in Science Teaching*, 58(9), 1359-1388.

[17] Böschl, F., Forbes, C., Lange-Schubert, K. (2023) Investigating scientific modeling practices in US and German elementary science classrooms: a comparative, cross-national video study. *Science Education*, 107(2), 368-400.

[18] Magana, A. J., Arigye, J., Udosen, A., Lyon, J. A., Joshi, P., Pienaar, E. (2024) Scaffolded team-based computational modeling and simulation projects for promoting representational competence and regulatory skills. *International Journal of STEM Education*, 11(1), 34.

[19] Webb, D. J., Paul, C. A. (2023) Attributing equity gaps to course structure in introductory physics. *Physical Review Physics Education Research*, 19(2), 020126.

[20] Sijmkens, E., De Laet, T., De Cock, M. (2026) Understanding and enhancing students' use of evaluation strategies during physics problem solving through reflective prompts. *Physical Review Physics Education Research*, 22(2), 020101.

[21] Wu, C. J., Liu, C. Y. (2021) Eye-movement study of high-and low-prior-knowledge students' scientific argumentations with multiple representations. *Physical Review Physics Education Research*, *17*(1), 010125.

[22] Hallström, J., Schönborn, K. J. (2019) Models and modelling for authentic STEM education: reinforcing the argument. *International Journal of STEM education*, 6(1), 22.