

Dual-diagnostic approach for jointly estimating students' conceptual understanding and misconceptions in forces and motion

Kuerbanjiang Jielili^{1,2}, Ziqi Yao^{1,2}, Shuhe Zhang^{1,2}, Mingpu Lan³, and Jianxin Yao^{1,2,*}

¹*School of Physics and Astronomy, Beijing Normal University, Beijing 100875, China*

²*Key Laboratory of Multiscale Spin Physics (Ministry of Education), Beijing Normal University, Beijing 100875, China*

³*Dongguan High School, Dongguan, Guangdong 523000, China*



(Received 26 September 2025; accepted 24 February 2026; published 9 April 2026)

Diagnosing students' mastery of physics concepts alongside their misconceptions is important for guiding effective physics instruction. Prior studies have analyzed distractor choices to identify misconceptions, and recent research has applied cognitive diagnostic modeling to obtain more fine-grained insights into student understanding. Building on this progress, we first proposed the two-stage multiple-choice deterministic inputs, noisy "AND" gate with hierarchy (TS-MC-DINA-H) model, which estimates the probabilities of both conceptual mastery and misconceptions through option-level response coding. On the basis of this model, we developed a cognitive diagnostic assessment tool for forces and motion. An analysis of data from 492 high school students validated the assessment tool. Moreover, the diagnostic results revealed that students with identical total scores often exhibited different mastery-misconception profiles, thus suggesting that diagnostic information can complement traditional score-based evaluation. The key goal of this study is to provide a new approach for simultaneously diagnosing concept mastery and a wide range of misconceptions at the individual level. This dual diagnostic perspective provides a more comprehensive picture of student learning and offers practical implications for targeted instruction and personalized remediation.

DOI: [10.1103/hrx4-gpd2](https://doi.org/10.1103/hrx4-gpd2)

I. INTRODUCTION

"Forces and motion" is a fundamental topic in physics and a key disciplinary core idea in science education standards worldwide [1,2]. Students' mastery of forces and motion is crucial for explaining a wide range of everyday phenomena and for understanding more advanced topics in science and engineering. However, numerous studies have shown that students often hold misconceptions about forces and motion that persist even after instruction [3–8]. These incorrect intuitive ideas can hinder students' physics learning [9–11]. Therefore, the physics education community has been focusing on how to better assess and address students' misconceptions in this domain. However, despite decades of work with concept inventories, concurrently diagnosing students' concept mastery and a broad spectrum of specific misconceptions at the individual level remains challenging.

In response to this need, we propose the two-stage multiple-choice deterministic inputs, noisy "AND" gate

with hierarchy (TS-MC-DINA-H) model for the joint inference of individual students' concept mastery and misconception profiles from multiple-choice responses. On the basis of this framework, we developed and validated a multiple-choice conceptual test on forces and motion. The following sections of this paper describe our work. We begin with the theoretical basis for our cognitive diagnostic approach and then provide background on misconceptions about forces and motion. We then describe the instrument development process, including how the test items and the underlying attributes were constructed. Next, we outline the methods for data collection and model implementation. We then present the results of the cognitive diagnostic analysis and highlight what they reveal about students' understanding of forces and motion. Finally, we discuss the implications of these findings for physics education and conclude with reflections on the utility of the TS-MC-DINA-H model in educational research and practice.

II. THEORETICAL BACKGROUND AND PROPOSED MODEL

A. Cognitive diagnostic assessment

Cognitive diagnostic assessment (CDA) offers a framework to obtain individualized profiles by combining test items with analytical models that can infer the status of multiple attributes for each student [12,13]. An attribute is a

*Contact author: yaojianxin@bnu.edu.cn

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

fine-grained latent variable that represents a specific knowledge, skill, or cognitive process required to solve an item [12,13]. In CDA, each item is designed to measure one or multiple specific knowledge attributes, and students' response patterns are used to determine which attributes they have mastered [14]. The linking of items to attributes is formalized in a Q-matrix [14,15], which is a binary matrix that indicates which attributes are required for a correct response to each item. At the core of CDA are cognitive diagnostic models (CDMs), which analyze response data using a given Q-matrix to estimate students' attribute mastery [12,13,16]. As multidimensional models, CDMs go beyond a single total score; they classify learners with respect to a set of specific knowledge components or attributes [17]. Among commonly used CDMs, the deterministic inputs, noisy "AND" gate (DINA) model assumes a conjunctive relationship between attributes. Under this assumption, an examinee must master all attributes required by an item to have a high probability of responding correctly, with slip and guess parameters accounting for random errors [18]. The deterministic inputs, noisy "OR" gate (DINO) model, by contrast, assumes a disjunctive relationship, so that mastering any one of the required attributes is sufficient to yield a high probability of success [19]. The generalized DINA (G-DINA) model further extends this framework by representing the probability of a correct response as a function of main effects and interaction effects of the underlying attributes, and it includes DINA, DINO, and several other CDMs as special cases [20]. The DINA model, G-DINA model, and related variants have been widely applied in educational settings, including mathematics education, science education, and language learning [21,22].

Recently, researchers have explored adapting these models to diagnose misconceptions by treating misconceptions as additional attributes in the model [23]. For instance, the Bug-DINO model focuses on identifying the presence of specific "bugs" or misconceptions [23], whereas the simultaneously identifying skills and misconceptions (SISM) model combines a knowledge model with a misconception model to classify learners into categories such as "mastered concept but holds misconception" versus "not mastered and holds misconception" [24]. These prior works show that misconception attributes can be inferred from answer patterns.

B. Diagnostic assessments of forces and motion

Forces and motion have been extensively studied in physics education research, and prior studies have cataloged a wide array of persistent misconceptions among learners [25,26]. Even after traditional instruction, students often retain intuitive beliefs that conflict with the formal principles of Newton's laws. For example, a common inertia-related misconception is the idea that an object in motion will gradually lose its motion unless a force keeps

"pushing" it, which reflects the impetus theory of motion [3].

A common approach to assessing students' conceptual understanding of forces and motion is the use of concept inventories, such as the force concept inventory (FCI) [25] and the force and motion conceptual evaluation (FMCE) [27]. Using the classical test theory (CTT) paradigm, researchers have extensively analyzed incorrect responses to identify misconceptions [25,27–33]. Complementing this work, studies employing item response theory (IRT) have further refined population-level diagnostics and provided richer insights into patterns of student thinking [34–37].

More recently, cognitive diagnostic approaches have been applied to forces and motion to yield learner-level profiles of both concept mastery and specific misconceptions. For example, using FCI data, a hybrid CDM-IRT model treats misconceptions as diagnostic attributes and reports a continuous ability score alongside probabilities of particular misconceptions [38]. Purpose-built CDA for Newton's laws has aligned Q-matrices via expert review and fitted responses with the DINA model [39]. A sequential bug-G-DINA design covering six attributes related to forces and motion validated the Q-matrix and found high misconception prevalence with limited sex- and grade-related differences [40]. In addition, post hoc DINA analyses of the FCI and FMCE within an evidence-centered design workflow refined item coding, estimated slip and guess parameters, produced skill classification accuracy, and underscored the necessity of Q-matrix validation [41].

However, previous applications of CDMs to diagnose physics misconceptions have mostly been post hoc analyses of existing tests rather than purpose-built diagnostic assessments [38,41]. Some existing models require relatively complex model specifications that are difficult for typical educators to implement on their own [38]. Furthermore, some studies have focused on only a narrow subset of misconceptions [39] mainly because increasing the number of attributes in cognitive diagnostic models inflates the parameter space and can lead to identifiability issues or nonconvergence [42]. However, physics education research has documented a broad range of persistent misconceptions in forces and motion [5,28]. Accordingly, there is a continuing need for psychometrically grounded, classroom-ready diagnostic methods and instruments that simultaneously estimate individual students' concept mastery and a broad spectrum of specific misconceptions.

C. The TS-MC-DINA-H model

The joint diagnosis of students' concept mastery and misconceptions from single-tier multiple-choice (MC) data faces two main challenges. First, scoring answers only as correct or incorrect overlooks the distractor information needed to identify specific misconceptions. Second, when we retain option-level information, CDMs can end up with too many parameters and identifiability problems.

To address the first challenge, we follow de la Torre [43] by coding each option with an option-level Q-matrix so that both keys and distractors carry diagnostic information that maps to latent classes. To address the second, we follow Elbulok [44] and use the subskill-misconception structure to support a two-step diagnostic strategy, thereby improving identifiability while retaining option-level information. In addition, to capture more misconceptions at finer granularity, we adopt a two-step strategy that, first, estimates concept mastery and then, under the hierarchy, confirms misconception profiles. Combining these elements, we propose the TS-MC-DINA-H model, which retains option-level detail while remaining efficient and stable for student-level diagnosis.

At the core of the model is a hierarchy (the “H” component) that imposes exclusion and prerequisite relations between concept attributes (A) and misconception attributes (M) [44]. The mastery of a given concept attribute excludes possessing its corresponding misconception attribute. For example, mastering Newton’s third law excludes the belief that “the heavier object exerts the greater force.” This hierarchy explains the model’s two-stage rule: if a student meets the concept requirements, we do not assume that they also hold a conflicting misconception; if the requirements are not met, we use the student’s response to determine which misconception best fits.

Based on this hierarchy, TS-MC-DINA-H follows a two-stage diagnostic process. In stage 1, the model evaluates concept mastery under DINA framework by using an item-level Q-matrix, thereby adopting the principle that each MC item encodes a unique combination of attributes. Students who master all the attributes required by the key are expected to select the correct option with the slip and guess parameters [43]. When mastery is not met, stage 2 uses the selected distractor to determine specific misconception attributes because the distractors are intentionally aligned with the misconception states in the option-level Q-matrix. In this way, single-tier responses yield two-tier diagnostic information without the addition of follow-up questions.

Taken together, these elements distinguish TS-MC-DINA-H from prior CDM-based approaches. The model is coupled to a purpose-built forces and motion diagnostic instrument. In this instrument, items and options are explicitly aligned with a hierarchical concept-misconception framework. TS-MC-DINA-H integrates option-level concept-misconception coding and a two-stage diagnostic rule into a single coherent modeling framework. This design enables it to support joint diagnosis of concept mastery and a relatively large set of misconception attributes without causing excessive parameter growth, thus helping to maintain identifiability and stable estimation. The next section presents instrument development for a forces and motion diagnostic aligned with physics curriculum standards and illustrates how these principles are operationalized.

III. METHODS

The workflow used in this study is summarized in Fig. 1. First, we developed the instrument by defining a concept attribute space (A-space) and a misconception space (M-space) and explicitly matched each misconception to its corresponding attribute. Afterward, we aligned each item’s key to the required concepts and each distractor to a specific misconception through an option-level Q-matrix. Second, we administered the instrument to high school students in classroom settings and coded responses at the option level. Third, we implemented the TS-MC-DINA-H model in two stages. Stage 1 estimated student mastery of each attribute by using an item-level concept Q-matrix under a DINA framework; conditional on nonmastery, stage 2 estimated the probabilities of holding the relevant misconceptions by using option-level submatrices. The outputs were compiled into student-level profiles that integrate total score, concept mastery, and misconception probabilities.

A. Instrument development

The development of the forces and motion diagnostic test was guided by both the content requirements of the high

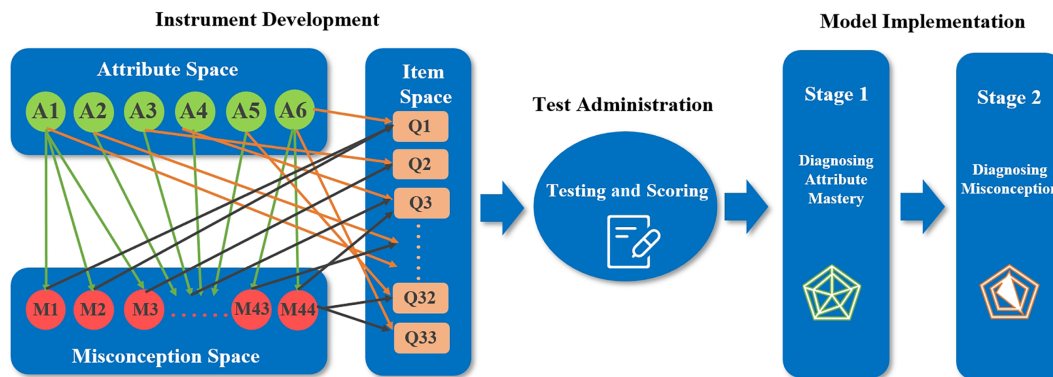


FIG. 1. Steps for the TS-MC-DINA-H diagnostic assessment.

school physics curriculum and the structural requirements of cognitive diagnostic modeling. The goal was to create a set of multiple-choice items that comprehensively cover forces and motion concepts, with each item capable of revealing specific concept mastery or misconception information. We followed a systematic design process similar to that used by an electrostatics diagnostic instrument development [45] but adapted here to the domain of forces and motion.

1. Concept attribute specification

We first constructed a concept attribute space (A-space) by identifying the key knowledge concepts in the domain of forces and motion. The A-space is based on curriculum standards and the literature on student learning in forces and motion. First, five attributes were tentatively identified. Five high school physics teachers were subsequently invited to review these attributes. After thorough discussion and refinement, six cognitive attributes were finally determined, as detailed in Table I.

On the basis of this optimized attribute space, the cognitive dependencies among the six attributes were articulated to produce the cognitive model shown in Fig. 2. The structure clarifies how an understanding of forces and motion is constructed. A1 and A2 serve as co-foundational attributes; A2 functions as a conceptual prerequisite for force-based reasoning in A3–A6. A1 focuses on motion (position, velocity, and acceleration)

without causes, whereas A3 treats equilibria where forces are analyzed in static configurations. A4 occupies the central node by formally coupling force and motion and mediates between descriptive and explanatory accounts. A5 and A6 represent advanced applications that presuppose a fluent command of A4, thereby illustrating the transfer of the force-motion relationship to specialized contexts. This configuration clarifies the logical progression among attributes, strengthens the theoretical coherence of the cognitive model, and provides a concrete basis for Q-matrix specification and diagnostic item development.

2. Identifying misconceptions

In parallel, we constructed a misconception space (M-space) by systematically reviewing prior research on common misunderstandings in the domain of forces and motion. The review identified 44 prevalent misconceptions related to forces and motion (Table II). These misconceptions were mapped to the six previously defined cognitive attributes, thus providing the basis for subsequent Q-matrix specification and item design.

3. Item design and alignment

With the concepts and misconceptions defined, we developed the item space, which consists of a pool of multiple-choice items. Each item was designed with one correct answer that requires the mastery of one or more concept attributes and three distractors, each intentionally

TABLE I. Optimized cognitive attributes of “forces and motion.”

Attribute code	Cognitive attribute name	Description of optimized cognitive attribute meaning
A1	Fundamentals of kinematics	Students can establish a particle model; choose an appropriate reference frame; and use physical quantities such as displacement, velocity, and acceleration to describe the linear (one-dimensional) motion state of objects and solve basic kinematics problems.
A2	Fundamentals of mechanics	Students understand that force is an interaction between objects (including contact and noncontact forces) and understand the concepts and properties of common forces such as gravity, elastic force, and friction.
A3	Statics	Students can carry out correct force analysis in real-world contexts and apply the properties of common forces together with vector addition and decomposition to solve problems involving the equilibrium of concurrent forces. They understand the concepts of resultant and component forces and the vector nature of force. They also understand that action and reaction act on different objects, and they do not confuse these with the balanced forces that act on a single object.
A4	Newton’s laws of motion	Students understand Newton’s first law (law of inertia) and second law and master the relationship between force and changes in motion (acceleration). They can apply Newton’s laws to analyze the forces and motion of objects.
A5	Projectile motion	Students understand the general laws of uniformly accelerated linear motion and master typical cases, such as free fall. They can analyze projectile motion by decomposing motion into horizontal and vertical components and extend the analysis to oblique projectile cases.
A6	Circular motion	Students understand both the conditions for uniform circular motion and the concepts and calculations of centripetal force and centripetal acceleration and can analyze and solve problems related to uniform circular motion.

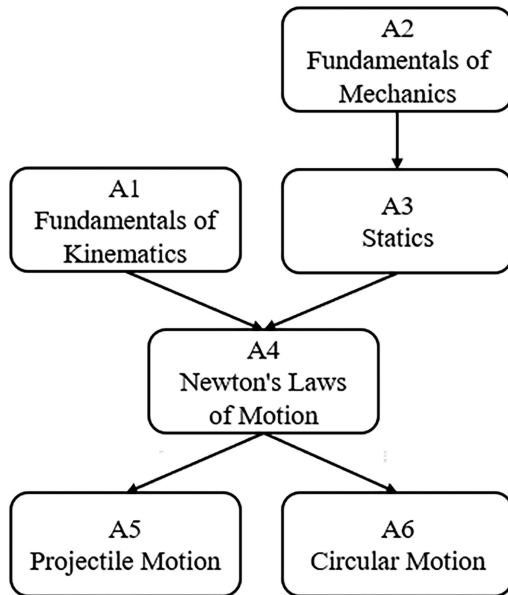


FIG. 2. Hierarchical cognitive model for forces and motion (A1–A6).

written to correspond to one of the misconception attributes. For example, item 10 describes a person pushing a heavy box with a constant horizontal force across a rough floor while the box moves at constant speed. The keyed option states that the push on the box and the friction acting on it are equal in magnitude and opposite in direction, thereby aligning with A3 (force analysis and equilibrium of concurrent forces) and A4 (linking zero acceleration to zero net force). The remaining options encode distinct misconceptions that constant motion direction implies a rightward net force (M29), that motion stops immediately once the sustaining force is removed because the “force runs out” (M28), and that friction on the box “cancels” the friction the box exerts on the floor by treating action-reaction forces as if they acted on the same object (M24).

While constructing the item space, to balance test length and identifiability, we ensured that each misconception attribute was assessed at least twice and that each concept attribute was assessed at least 4 times. This approach follows prior work emphasizing the use of multiple items per attribute when slipping and guessing parameters are estimated [60]. For conjunctive diagnostic models such as the DINA model, when both slipping and guessing parameters are unknown, each attribute must be required by at least three items for identifiability [42]. Using two items for each misconception attribute thus represents a practical compromise between these theoretical considerations and constraints on test length.

Each draft item underwent expert review for content validity. A focus group consisting of two high school physics teachers and one physics education researcher evaluated the items for scientific accuracy, clarity of wording, and alignment with the intended concept and

misconception attributes. Their feedback led us to revise several items by, for instance, rewording confusing stems, ensuring that distractors were plausible yet clearly tied to the target misconception, and verifying that no unintended cues would give away the correct answer [61]. After revising the multiple-choice items, we selected those that best covered the attribute list and performed well in qualitative evaluation, thereby narrowing the set to 33 items. The full original Chinese test and its English translation are provided in Supplemental Material [62].

4. Q-matrix construction

In parallel with item writing, we developed an extended option-level Q-matrix (options $\times$ attributes) that encodes alignment to both concept attributes (A1–A6) and misconception attributes (M1–M44). Concept columns indicate the knowledge required to select the keyed option; misconception columns indicate which distractor each misconception explains. Hierarchical prerequisites were represented in the concept columns using an implicit Q-matrix [63]. For any item mapped to a higher-level attribute, its prerequisite attributes were also coded as required for that item. It should be noted that this implicit-Q encoding does not strictly eliminate all theoretically impermissible mastery profiles because we did not impose a structured prior that assigns zero prior probability to those profiles [63]. Separately, the intended option-level exclusion rule is implemented by setting all misconception columns to zero for the keyed option and setting all concept columns to zero and the corresponding misconception column to one for each distractor; this ensures that mastery of the required concepts does not co-occur with the paired misconception.

To illustrate the construction procedure more concretely, we take item 10 as a fully worked example (see Fig. 3). The item reads as follows:

EXAMPLE QUESTION 10

A person pushes a heavy box to the right across a rough horizontal floor using a constant horizontal force F , and the box moves to the right at a constant speed. Regarding the forces acting on the box, which of the following statements is correct?

- The pushing force F exerted by the person on the box is equal in magnitude and opposite in direction to the frictional force exerted by the floor on the box.
- Since the box is moving to the right at a constant speed, the net force on it is also directed to the right.
- When the pushing force stops, the box will immediately stop moving because no force remains to keep it going.
- The fact that the box moves to the right at constant speed shows that the frictional force exerted by the floor on the box cancels the frictional force exerted by the box on the floor.

TABLE II. Catalog of misconceptions about forces and motion ($n = 44$) and their alignment to six cognitive attributes. Note that for each misconception, a single reference documenting this idea in the literature is cited. When the same misconception has been reported in multiple studies, an early, widely cited source is retained as the representative citation, and additional related work is discussed in the main text where appropriate.

Cognitive attribute	No.	Misconception	Source
A1. Kinematics basics	M1	Confusing distance with displacement.	Yao and He [46]
	M2	Believing acceleration exists only when the speed of an object increases, not decreases.	
	M3	Equating greater change in position with faster motion.	
	M4	Assuming identical positions imply identical velocities.	
	M5	Difficulty understanding negative velocity.	
	M6	Confusing instantaneous and average velocity.	
	M7	Overlooking directional changes as forms of acceleration.	
	M8	Believing velocity and acceleration are directly proportional.	
	M9	Assuming larger velocity changes always correspond to greater acceleration.	
A2. Mechanics basics	M10	Thinking acceleration cannot exist without velocity.	Kim and Pak [47]
	M11	Believing force is caused by only living or active objects.	Halloun and Hestenes [6]
	M12	Thinking force requires physical contact.	
	M13	Viewing force as a property of a single object rather than an interaction between two objects.	Reiner <i>et al.</i> [48]
	M14	Assuming friction always opposes motion.	Kızılcık <i>et al.</i> [49]
	M15	Viewing force as a property or quantity that an object can possess and exert based on its weight or motion.	Brown [50]
	M16	Assuming the center of mass must lie within the object.	Klammer [51]
	M17	Believing gravity is uniform everywhere on Earth.	Gönen [52]
	M18	Misunderstanding the relationship between elastic (restoring) force and deformation.	Yao and He [46]
A3. Statics	M19	Equating weightlessness with the absence of gravity.	Gönen [52]
	M20	Believing gravity increases with distance.	Sari <i>et al.</i> [53]
	M21	Assuming normal force always equals gravitational force.	
	M22	Confusing action-reaction pairs, balanced forces, and the resultant.	
	M23	Believing only “active” forces exist.	Kruger <i>et al.</i> [54]
	M24	Thinking action-reaction forces can cancel out.	Bayraktar [55]
	M25	Believing objects move only when action force exceeds reaction force.	Yao and He [46]
	M26	Assuming larger objects exert greater forces.	Hestenes <i>et al.</i> [25]
	M27	Believing acceleration depends only on pulling forces.	Yao and He [46]
A4. Newton’s laws of motion	M28	Thinking force is needed to sustain motion.	Clement [3]
	M29	Assuming net force direction must align with motion direction.	Halloun and Hestenes [6]
	M30	Viewing rest as the “natural” state of objects.	
	M31	Assuming faster objects experience greater forces.	
	M32	Believing objects contain an internal “force of motion.”	Bayraktar [55]
	M33	Thinking objects stop when their “force” is depleted.	Halloun and Hestenes [6]
	M34	Assuming trajectory aligns with net force direction.	
	M35	Believing that once an object is released, it loses its prior horizontal motion and moves only straight down under gravity.	Clement [3]
	M36	Thinking heavier objects fall faster.	Caramazza <i>et al.</i> [56]
A5. Projectile motion	M37	Assuming a horizontal force sustains projectile motion.	Gunstone and White [4]
	M38	Believing the net force always points along the path.	Prescott and Mitchelmore [57]
	M39	Thinking force is zero at the peak of vertical motion.	Neidorf <i>et al.</i> [58]
	M40	Confusing “uniform speed” with “uniform velocity.”	Mutsvangwa [59]
	M41	Believing centrifugal force exists as a real outward force.	
	M42	Assuming a tangential force is required to keep an object in uniform circular motion.	Clement [3]
	M43	Thinking constant speed implies constant acceleration, thus ignoring its continuously changing direction.	Mutsvangwa [59]
	M44	Believing objects continue along curved paths after forces are removed.	McCloskey <i>et al.</i> [7]

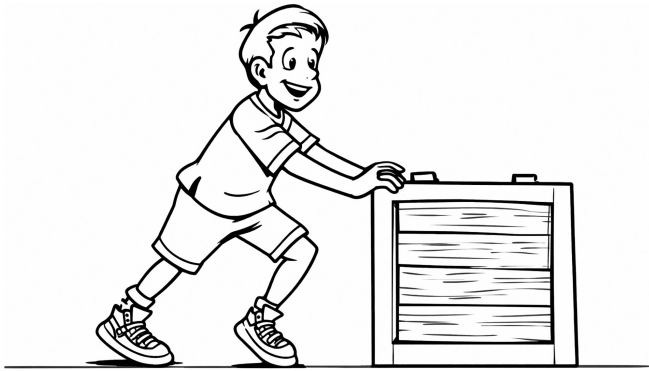


FIG. 3. Figure for Q10.

Correct understanding of this situation requires students to recognize that the box is in a state of dynamic equilibrium: a constant velocity implies zero net force, and the horizontal contact force from the person must be balanced by the frictional force from the floor. In our cognitive model, these ideas are represented by attributes A3 and A4. Because A1 and A2 are hierarchical prerequisites for A3 and A4, the keyed option (option A) is coded with the concept vector (1, 1, 1, 1, 0, 0), and all misconception columns are set to 0.

Each distractor, by contrast, encodes a specific misconception and is therefore not linked to any concept attributes. Option B states that “since the box is moving to the right at a constant speed, the net force on it is also directed to the right,” which reflects the misconception that constant motion requires a net force in the direction of motion; this is represented by misconception attribute M29. Option C claims that “when the pushing force stops, the box will immediately stop moving because no force remains to keep it going,” which corresponds to an impetus-style

misconception that a continuous force is needed to maintain motion (attribute M28). Option D asserts that the frictional force exerted by the floor on the box cancels the frictional force exerted by the box on the floor, revealing a confusion between Newton’s third law and the net force on a single object (attribute M24). In the option-level Q-matrix, the rows for Options B, C, and D therefore have zeros in all concept columns and a single 1 in the M29, M28, and M24 columns, respectively, with all other misconception columns set to 0.

Figure 4 visualizes this encoding for item 10 as an option-level Q-matrix heatmap. Each row corresponds to one option (Q10_A–Q10_D), and each column to a concept attribute (A1–A6) or a misconception attribute (M1–M44). Blue cells indicate concept attributes coded as 1, red cells indicate misconception attributes coded as 1, and white cells indicate 0. The numbers inside the cells show the underlying binary coding. As can be seen, the keyed option Q10_A is associated only with the required concept attributes A1–A4, whereas each distractor is associated with exactly one misconception attribute, providing a clear blueprint for subsequent diagnostic analysis.

B. Test administration

1. Participants

In this study, participants were drawn from multiple secondary schools in three provinces in China (Shandong, Guizhou, and Xinjiang), which represent a mix of urban and suburban settings. A total of $N = 496$ students (approximately 15–16 years old) participated in the test. The sample included a roughly equal proportion of male and female students and represented a range of achievement levels in science. Tests were administered during regular class periods under teacher supervision.

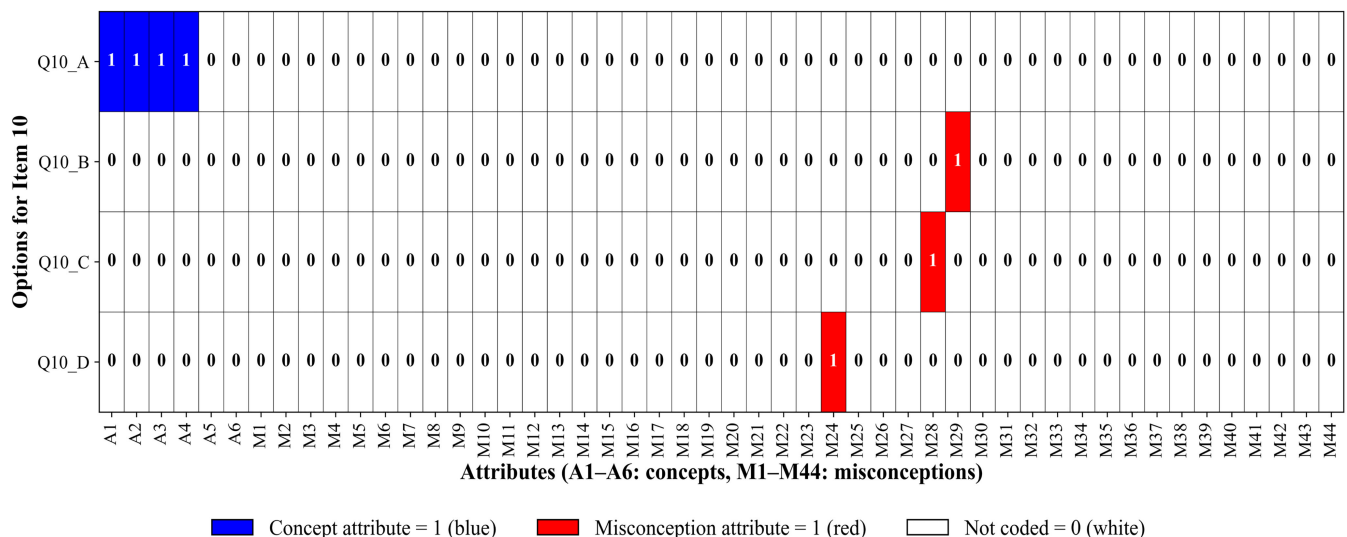


FIG. 4. Option-level Q-matrix for item 10.

The diagnostic test was given in a paper-and-pencil format with a 60-min time limit, which was sufficient for almost all students to complete all 33 items. Students were instructed to choose the best answer for each question. No calculators or external aids were permitted. To encourage honest responses (which are important for diagnostic validity), we emphasized to the students that the test scores would not count toward their grades and that the purpose was to understand their thinking to help improve teaching. After administration, the answer sheets were collected and recorded for initial analysis as described below.

2. Data coding and scoring for the diagnostic model

In preparation for the cognitive diagnostic analysis, students' responses were not simply coded as "correct" or "incorrect" but were represented at the option level. Specifically, the response matrix was constructed such that each row corresponded to a student ID, and each column represented one option of an item. If a student selected a particular option, the corresponding cell was coded as "1" and "0" otherwise. This option-level encoding allowed us to directly link incorrect responses to the misconception attributes they represent. For example, if a student selected distractor (C) on item 5, which was designed to capture the misconception of "believing only active forces exist," the coding scheme explicitly preserved this evidence. In this way, the matrix captured multivariate evidence from each item. Correct options provided evidence of concept mastery, while each distractor indicated a distinct misconception.

C. Model implementation

We applied the TS-MC-DINA-H model to the collected data in R by using the "GDINA" package for CDM estimation [64], the "CDM" package for auxiliary diagnostics (e.g., classification indices) [65], the "LTM" package for 2PL IRT (Two-Parameter Logistic Item Response Theory) [66], and the "PSYCH" package for reliability and classical indices [67]. First, we imported the extended, option-level Q-matrix and the student response file. From the extended matrix, we derived an item-level concept Q-matrix (items $\times$ concept attributes) by keeping the keyed-option rows and retaining their A1–A6 columns at the item level; distractor rows were retained for the later misconception analysis. Student responses were reshaped to a person $\times$ item matrix and scored dichotomously (1 = keyed option chosen; 0 otherwise).

To examine instrument quality and reliability, we reported classical test indices including difficulty (proportion correct), discrimination (item-total point-biserial), and Cronbach's α [67]. We also fitted a unidimensional 2PL model to obtain item discrimination, item difficulty, marginal reliability, test information, and pointwise reliability across θ [66].

Within the CDM framework, we fit DINA, DINO, and G-DINA on the concept Q-matrix to identify the best-fitting model. We compared models by using $-2 \log$ -likelihood, AIC, and BIC and evaluated absolute fit with limited-information statistics (M2, RMSEA₂, and SRMSR) [68,69]. From the selected model, we reported person-level posterior mastery probabilities and maximum likelihood estimation (MLE) mastery profiles for A1–A6, together with attribute-level classification accuracy and consistency; we also summarized the corresponding classification outputs available in the "CDM" package [65,70].

Diagnosis proceeded in two hierarchy-aware stages aligned with our instrument. Stage 1 estimated concept mastery (posterior probabilities and mastery profiles) on the concept Q-matrix. Stage 2 proceeded attribute by attribute: for each concept A_k , we focused on students not yet classified as mastering A_k ; constructed an option-level misconception matrix by linking the distractor options for items targeting A_k to the misconception tags designed for that attribute; and fitted separate DINA models to obtain person-level posterior probabilities and classifications for those misconceptions. Although misconception estimates were obtained from separate submodels rather than from a single joint model with strict constraints, conditional on stage 1 mastery, we reported and interpreted these estimates to respect the intended hierarchy. This approach allowed us to implement the proposed model by using existing R packages rather than developing new R code.

IV. RESULTS

A. Item statistics and reliability

Overall, the 33-item diagnostic test on forces and motion showed strong psychometric properties. Table III summarizes these item-level results by reporting, for each item, CTT difficulty (p) and point-biserial discrimination alongside 2PL parameters (discrimination a and difficulty b), thereby allowing readers to inspect targeting and discrimination across the test against those conventional guidelines. Under classical test theory (CTT), most items demonstrated moderate difficulty (the proportion of correct responses ranged from 0.50 to 0.85) and high discrimination (the point-biserial correlations were approximately 0.40 $\sim$ 0.75); consequently, these items fall squarely within widely cited screening recommendations for well-functioning items (difficulty centered between approximately 0.30 and 0.80; the point-biserial correlations ≥ 0.30 were considered good, whereas 0.20 was a common minimum [71,72]). Only a few items were very difficult (e.g., item 8, $p \approx 0.20$), and a small number showed weaker discrimination. Under IRT (2PL), most items showed strong discrimination, with a value around 1.6 $\sim$ 2.2 (0.65 $\sim$ 1.34 = moderate; 1.35 $\sim$ 1.69 = high; ≥ 1.70 = very high [73]), and difficulties largely between -1.1 and -0.5 . Overall, the test was relatively easy for the

TABLE III. Item-level CTT (p, point-biserial) and 2PL (a, b) statistics for all 33 items.

Item	Difficulty (CTT)	Discrimination (CTT)	Difficulty (IRT)	Discrimination (IRT)
Q1	0.736	0.624	-1.03473	2.174128
Q2	0.631	0.687	-0.63101	2.000898
Q3	0.857	0.434	-1.72877	1.610368
Q4	0.645	0.647	-0.69488	1.807932
Q5	0.823	0.587	-1.29851	3.525136
Q6	0.635	0.608	-0.67309	1.547382
Q7	0.748	0.553	-1.14917	1.658293
Q8	0.202	0.13	6.423789	0.214366
Q9	0.591	0.749	-0.47224	2.458813
Q10	0.744	0.615	-1.07541	2.067372
Q11	0.673	0.648	-0.80072	1.948706
Q12	0.732	0.662	-1.00468	2.492246
Q13	0.774	0.413	-1.65142	0.956645
Q14	0.722	0.573	-1.03473	1.643851
Q15	0.52	0.276	-0.26009	0.473249
Q16	0.831	0.461	-1.57615	1.606107
Q17	0.619	0.73	-0.58055	2.318026
Q18	0.722	0.652	-0.98042	2.15831
Q19	0.565	0.596	-0.38089	1.410589
Q20	0.661	0.715	-0.7423	2.327592
Q21	0.53	0.591	-0.23309	1.387
Q22	0.79	0.579	-1.24838	2.235292
Q23	0.75	0.666	-1.06541	2.705184
Q24	0.762	0.548	-1.22554	1.611144
Q25	0.52	0.637	-0.18928	1.604369
Q26	0.72	0.603	-1.00252	1.807453
Q27	0.597	0.653	-0.50344	1.694328
Q28	0.667	0.617	-0.79532	1.688492
Q29	0.607	0.63	-0.55204	1.563704
Q30	0.659	0.699	-0.73576	2.254505
Q31	0.46	0.511	0.101454	1.09265
Q32	0.534	0.483	-0.26617	1.006522
Q33	0.593	0.66	-0.48326	1.794545

present sample. However, this aligns with our primary goal of diagnosing basic concepts and misconceptions, which are especially prevalent among lower-performing students.

The test also demonstrated high internal consistency, with a Cronbach's α of 0.94, which is typically interpreted as excellent for high-stakes or diagnostic uses [74,75]. Consistent with this, a 2PL IRT analysis indicated that the test provides substantial information over a wide span of ability. The test information function peaks at a level implying very high conditional precision (approximately 0.99 at the peak) [73,76]. In addition, the IRT-based marginal reliability of the test scores was 0.884. Marginal reliability is a single-number summary of test precision in IRT—analogue to classical reliability [76,77]. In practical terms, this means that the instrument can very reliably rank-order students' understanding of forces and motion.

Moreover, the cognitive diagnostic classifications were stable. The model's classification consistency for each

TABLE IV. Attribute-level diagnostic classification consistency and accuracy indices.

	Classification accuracy	Classification consistency
MLE_A1	0.993	0.986
MLE_A2	0.955	0.914
MLE_A3	0.773	0.988
MLE_A4	0.774	0.998
MLE_A5	0.762	0.982
MLE_A6	0.719	0.912

attribute was high (e.g., the reliability of mastery decisions for attributes A1–A2 exceeded 0.90 and was approximately 0.80 for A3–A6). We computed the diagnostic classification indices at the attribute level based on established formulations and their R implementation [64,70]. These indices are reported in Table IV.

B. Model fit and selection

We compared several cognitive diagnostic models to ensure the best fit for our data. Table V summarizes the key indices used for relative and absolute fit evaluation. The DINA model provided the best balance between fit and parsimony, thereby yielding the lowest information criterion (AIC = 15,079; BIC = 15,622); this result is consistent with the information-theoretic view that a smaller AIC/BIC indicates a better trade-off between fit and complexity [78–80]. The DINO model produced a slightly higher AIC/BIC (AIC = 15,095; BIC = 15,637). The G-DINA model attained a higher log-likelihood but with many additional parameters, thereby producing a substantially larger AIC (15,223) and BIC (17,330). This finding is expected because G-DINA is a saturated framework encompassing reduced CDMs, so penalized criteria often favor simpler models when added flexibility risks overfitting [20].

Absolute fit was good for all the models (e.g., $RMSEA_2 < 0.05$), and the limited-information statistics further documented fit at the test level. G-DINA yielded a lower $RMSEA_2$ (0.025) than did DINA (0.030), thereby reflecting a closer reproduction of observed patterns [68,69,81,82]. Given DINA's superior AIC/BIC and its parsimony relative to G-DINA, we selected the DINA framework as the basis for subsequent diagnostic modeling. All follow-up analyses employed our extended DINA-based model (TS-MC-DINA-H), which preserves DINA's noisy-AND structure while incorporating misconceptions hierarchically within the diagnostic workflow.

In addition to evaluating model fit for the stage 1 concept-diagnostic models, we also examined the fit of the stage 2 misconception submodels. Stage 2 proceeded attribute by attribute and fitted separate DINA-based submodels to the misconception option sets linked to each concept attribute. Table VI reports relative fit indices and

TABLE V. Goodness-of-fit indices for alternative cognitive diagnostic models.

Model	−2LL	AIC	BIC	M2	RMSEA ₂	SRMSR
DINA	7410.522	15,079.04	15,621.69	627.1382	0.0302	0.0496
DINO	7418.395	15,094.79	15,637.44	631.6604	0.0305	0.0496
GDINA	7110.501	15,223	17,330.5	78.561	0.025	0.0406

TABLE VI. Goodness-of-fit indices for stage-2 misconception submodels (DINA).

Submodel	logLik	AIC	BIC	M2	RMSEA ₂	SRMSR
Stage 2_A1	−2758.18	7646.364	12,126.37	533.3725	0.0606	0.0825
Stage 2_A2	−3194.37	8514.746	12,986.34	610.4144	0.0723	0.1000
Stage 2_A3	−2367.53	4921.067	5312.279	189.7026	0.0472	0.0704
Stage 2_A4	−2129.70	4581.394	5258.653	236.8219	0.0447	0.0681
Stage 2_A5	−1969.81	4117.619	4492.074	160.9295	0.0545	0.0667
Stage 2_A6	−2018.71	4143.422	4366.371	136.6686	0.0652	0.0699

limited information absolute fit statistics for each stage 2 submodel. As descriptive reference values, RMSEA values below 0.05 are often interpreted as close fit and values between 0.05 and 0.08 as acceptable fit, whereas values above 0.10 may indicate poor fit [83]. We use these RMSEA reference values as descriptive guidelines for interpreting RMSEA₂. Under this interpretation, the stage 2 submodels showed acceptable absolute fit (RMSEA₂ = 0.045–0.072; SRMSR = 0.067–0.100), supporting the adequacy of the second-stage estimation and the resulting misconception probability profiles.

C. Attribute mastery profiles

Using the selected model, we estimated each student's probability of mastering six key physics concepts. The average mastery probabilities by attribute, which reveal both strengths and weaknesses in the cohort, are shown in Fig. 5. Students showed the greatest mastery of statics (A3,

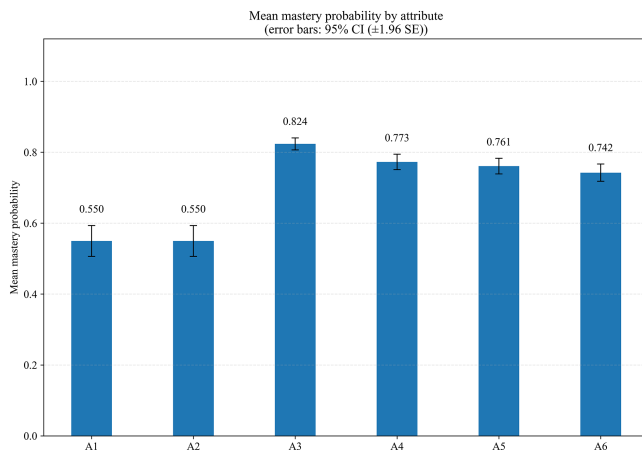


FIG. 5. Mean mastery probability for each cognitive attribute across all students.

≈0.824) and Newton's laws of motion (A4, ≈0.773); their mastery of projectile motion (A5, ≈0.761) and circular motion (A6, ≈0.742) was also relatively strong. In contrast, their mastery of basic kinematics (A1) and basic force concepts (A2) was much weaker (both ≈0.55). In other words, just over half of the kinematics and basic force content was mastered by the average student, whereas three-quarters or more of the statics and Newton's laws content was mastered. One plausible explanation is that these foundational attributes may be strongly associated with prominent misconceptions that are triggered by everyday experience and intuitive ways of making sense of physical phenomena [84,85]. Given that our instrument is single-tier multiple-choice and each distractor is intentionally written to represent a specific misconception, students who hold an intuitive misconception may instinctively select the corresponding distractor even if they have been exposed to the formal definition. This mechanism may depress the estimated mastery probabilities for A1 and A2, where misconceptions about basic kinematic quantities and basic force ideas are both numerous and cognitively compelling [3,6,25]. Such robust prior conceptions can persist and constrain the interpretation of instruction, making foundational understanding especially resistant to change [86,87]. In contrast, items targeting A3–A6 are closer to well-practiced classroom problem types, where students may achieve higher success through repeated practice and more stable procedures [88]. In these more procedural contexts, misconception-driven option attraction may be less dominant.

The distribution of mastery profiles was also informative. The frequency of the 10 most common attribute mastery profiles in the sample is shown in Fig. 6. With each student being represented as a binary vector (1 = mastered, 0 = not), the most frequent profile was 111111 (mastery of all six attributes), which was observed in 223 students, accounting for over half of the sample.

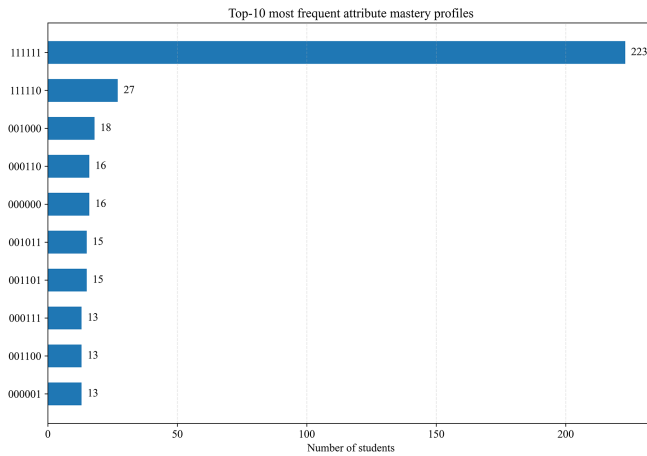


FIG. 6. Frequency of the 10 most common attribute mastery profiles in the sample.

The next most common pattern was 111110 (all but A6, circular motion), observed in 27 students, and the pattern 000000 (no attributes mastered), observed in 16 students. Several midlevel patterns also appeared, for example, 000100 (only A3), 001001 (A3 and A6), and 000110 (A4 and A5), indicating isolated strengths and gaps across learners.

D. Misconception classification outcomes

For students who had not mastered an attribute, the model estimated their probabilities of holding each related misconception. By averaging these probabilities across all the students who did not master a given attribute, we obtained a profile of the most prevalent misconceptions associated with each content area. These findings are presented in Fig. 7. In summary:

A1 (Kinematics)—Among the 220 students who did not master basic kinematics (A1), the two most prevalent misconceptions were M1 (confusing distance with displacement) and M8 (believing velocity and acceleration are directly proportional), with mean estimated posterior probabilities of approximately 0.69 and 0.66, respectively. Other kinematics-related misconceptions, such as M10 (the belief that acceleration cannot exist without velocity; $p \approx 0.39$), were also present but less common. Overall, students lacking A1 showed significant difficulty in distinguishing between closely related concepts (e.g., distance versus displacement, velocity versus acceleration) and a notably poor understanding of acceleration itself.

A2 (Basic forces)—Among the 194 students who did not master fundamental force concepts (A2), the most prevalent misconception was M15 (viewing force as a property or quantity that an object can possess and exert based on its weight or motion), with a mean estimated probability of 0.705. Two other notable misconceptions among these students were M14 (the

view that friction always opposes motion; $p \approx 0.47$) and M18 (misunderstanding the relationship between elastic force and deformation; $p \approx 0.45$).

A3 (Statics)—Among the 114 students who had not mastered static equilibrium and interaction forces, the most common misconceptions were M22 (confusing action-reaction pairs with balanced forces; $p \approx 0.80$) and M21 (assuming normal force always equals gravitational force; $p \approx 0.71$). A similarly prevalent misconception was M26 (assuming larger objects exert greater forces; $p \approx 0.69$). Other observed misconceptions within this group included M25 ($p \approx 0.60$), M24 ($p \approx 0.40$), and M23 ($p \approx 0.37$).

A4 (Newton’s laws of motion)—Among the 116 students who had not mastered Newton’s laws, the top misconceptions were M31 (assuming faster objects experience greater forces; $p \approx 0.68$), M32 (believing objects contain an internal “force of motion”; $p \approx 0.58$), and M29 (assuming net force direction must align with motion direction; $p \approx 0.55$). Many also believed M28 (thinking force is needed to sustain motion; $p \approx 0.47$), with fewer believing M30 (viewing rest as the “natural” state of objects; $p \approx 0.30$), M33 (thinking objects stop when their “force” is depleted; $p \approx 0.26$), and M27 (believing acceleration depends only on pulling forces; $p \approx 0.14$).

A5 (Projectile motion)—Among the 136 students who had not mastered two-dimensional motion, the most common errors were M38 (the net force always points along the path; $p \approx 0.62$), M37 (a constant horizontal force keeps a projectile moving forward; $p \approx 0.55$), and M39 (thinking force is zero at the peak of vertical motion; $p \approx 0.54$). Fewer chose M35 ($p \approx 0.37$), M34 ($p \approx 0.24$), and M36 ($p \approx 0.21$).

A6 (Circular motion)—Among the 150 students who had not mastered circular motion, the most common misconception was M41 (believing centrifugal force exists as a real outward force; $p \approx 0.53$). The next common misconceptions were M42 (assuming a tangential force is required to keep an object in uniform circular motion; $p \approx 0.40$) and M43 (constant speed means constant acceleration; $p \approx 0.29$), which ignores the acceleration direction changing toward the center. A notable minority also believed M44 (believing objects continue along curved paths after forces are removed; $p \approx 0.25$) and M40 (uniform speed means uniform velocity; $p \approx 0.25$).

E. Illustrative individual profiles

For every student, we used our TS-MC-DINA-H model to estimate the mastery probability for each attribute (A1 ~ A6) and the belief probability for each misconception (M1 ~ M44). We do not list every student’s profile here. Instead, highlighting how our approach is more informative than a single total score, we focus on two

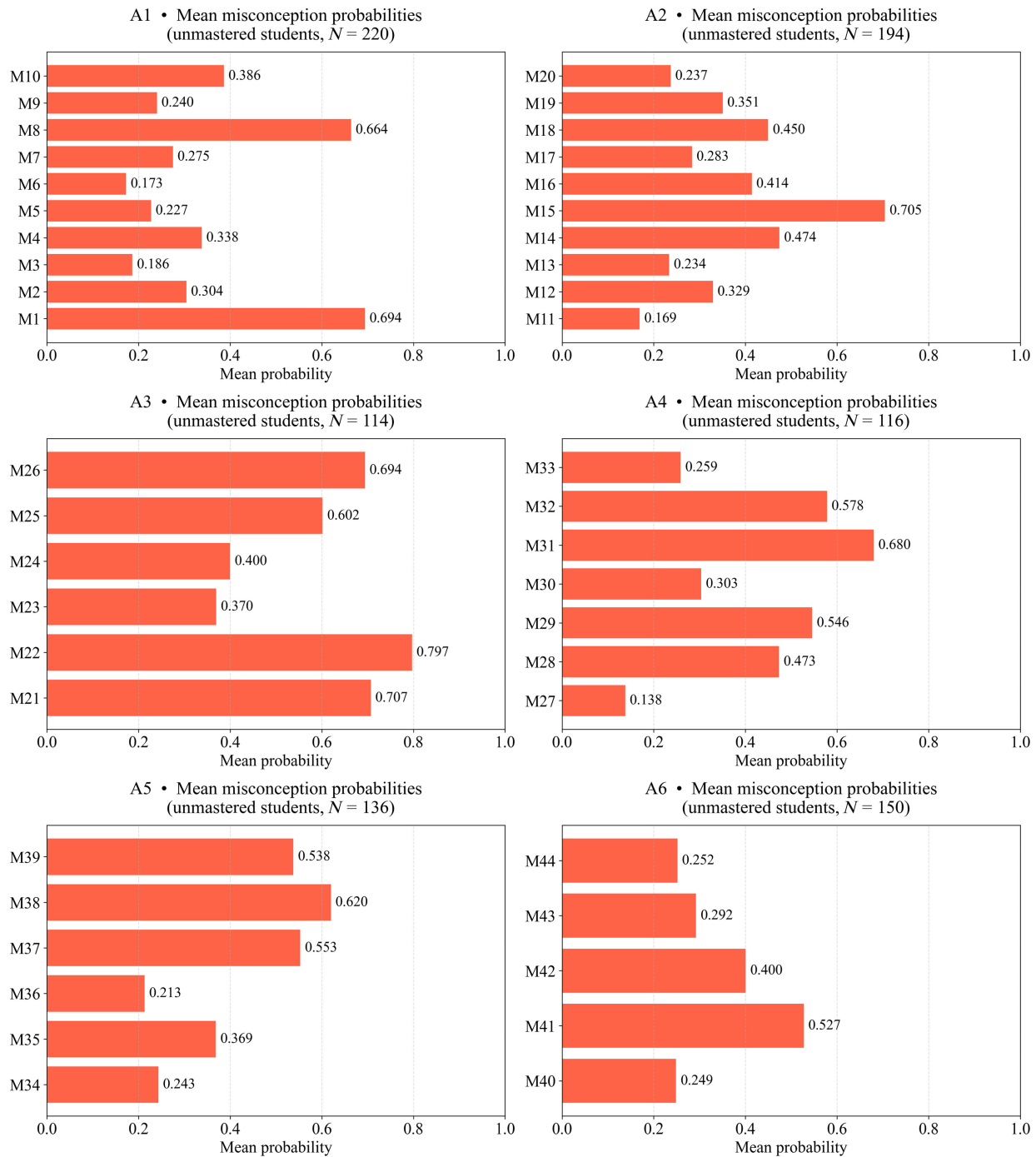


FIG. 7. Mean misconception probabilities for nonmastery subgroups by attribute.

types of comparisons: (1) students with the same total score but different attribute mastery patterns and (2) students with the same attribute mastery pattern but different misconception profiles.

Figure 8 compares radar charts of attribute mastery probabilities for five example pairs of students with equal total test scores. Each radar plot corresponds to a pair of students (denoted by blue and orange) who earned the same total score on the 33-item test but who exhibited different

levels of mastery of the six attributes (A1 ~ A6). For example, as shown in the radar chart, student 108 (blue) shows uniformly high mastery probabilities across attributes (approximately $0.87 \sim 0.95$), whereas student 132 (orange) remains low on most attributes (approximately $0.25 \sim 0.51$). The students' binary mastery profiles also differ (108: 101110; 132: 110100). Although the students have identical total scores, student 108 answered many multiattribute items leading the model to assign high

mastery probabilities across attributes. In contrast, student 132's correct answers concerned mostly single-attribute items and fewer diagnostic items and included errors on key multiattribute items (especially those involving A3 and its combinations), thus yielding lower posterior mastery probabilities. Thus, identical total scores can arise from very different patterns of evidence, which CDM captures but a sum score cannot. This example illustrates that two students can arrive at the same total score by very different paths. More broadly, across the four pairs shown in Fig. 8, we consistently observe that “same score” does not imply “same knowledge state.” For every pair of students with equal scores that we examined, their radar charts of attribute mastery differ in shape—sometimes dramatically.

Figure 9 shows heatmaps of misconception probabilities for four example pairs of students with identical attribute mastery profiles. Each pair of columns shows two students

who have the same set of mastered attributes; the heatmaps show the model-estimated probability that each student holds each misconception (M1 ~ M44). Blue indicates a lower probability, red indicates a higher probability, and white indicates a midlevel probability.

We can see that four pairs of students share the same attribute-mastery profile but display distinct misconception landscapes. For example, in the first pair (profile 000000; 154 versus 291), student 154 generally has lower misconception probabilities, whereas student 291 has many higher-probability misconceptions. Across all panels, the heatmaps reveal numerous item-level contrasts. In these cases, one student shows a high misconception probability for an item while another shows a low probability. This underscores that identical attribute profiles do not imply identical misconceptions, and misconception diagnostics provide information beyond a single score or profile.



FIG. 8. Radar chart comparisons of attribute mastery probabilities for five example pairs of students with equal total test scores.

Four Pairs: Misconception Probability Heatmaps (Same Attribute Profile)

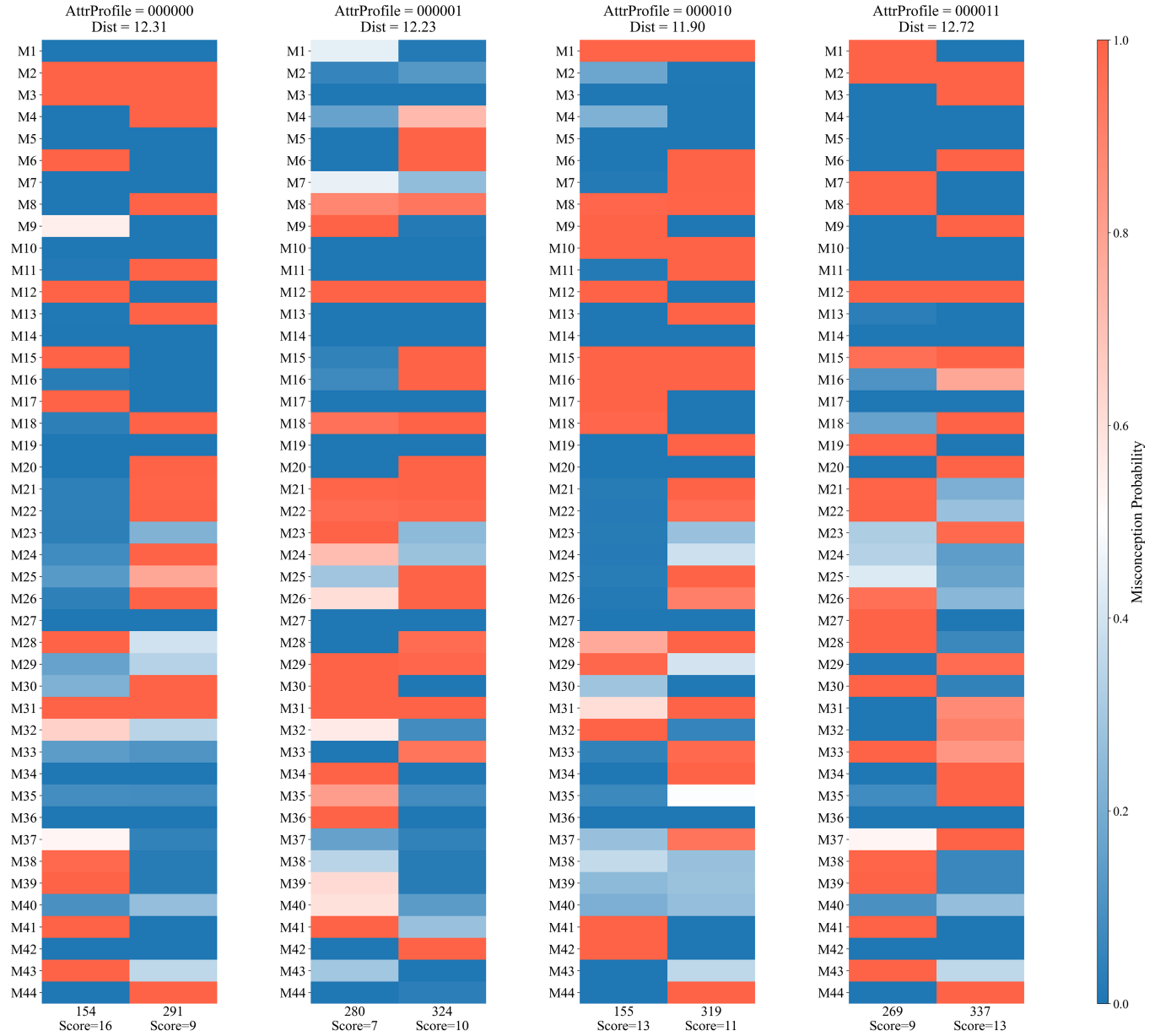


FIG. 9. Heatmaps of misconception probabilities for four example pairs of students with identical attribute mastery profiles.

V. DISCUSSION

In this study, we proposed the TS-MC-DINA-H model and developed and validated a cognitively diagnostic assessment for forces and motion. The results show that the instrument is psychometrically sound, thus supporting its confident use for individual diagnostic profiling within the cognitive diagnosis modeling framework [13]. In addition, the assessment quantifies long-documented misconceptions at the individual level. Many students still endorse impetus-like beliefs and struggle with Newton's third law and normal forces—patterns long noted in Physics Education Research (PER) [6,25–27]. In addition,

equal total scores can hide very different cognitive profiles: students with the same total score differ in mastered concepts and in the misconceptions they hold, thus echoing findings in other domains [21,89]. This point is also reflected in the attribute-level results. In particular, we observed a seemingly counterintuitive pattern that the prerequisite attributes A1 and A2 show lower mean mastery than A3 and A4, which provides a useful example of why a dual focus on mastery and misconceptions matters for interpretation.

To interpret this pattern, it is important to note that it should not be interpreted as students “skipping”

prerequisites in the cognitive hierarchy. The attribute hierarchy encodes prerequisite relations for expertlike reasoning, whereas posterior mastery probabilities are context-dependent estimates of the likelihood that students can successfully apply the targeted knowledge under the assessment conditions [13]. Thus, relatively high mastery for A3/A4 can coexist with weak A1/A2. This may occur because students rely on practiced procedures or context-specific strategies in typical force-analysis tasks, even when their foundational representations are fragile and vulnerable to misconception activation [88]. This interpretation is also consistent with the diagnostic logic that students who endorse a particular misconception may be more likely to select the corresponding distractor [90], so foundational weaknesses may be overlooked if one considers only higher-level task performance. Our approach helps make such hidden weaknesses more visible by jointly estimating concept mastery and misconception endorsement from option-level responses. At the same time, these results also prompt a deeper question for future work: To what extent, and through what mechanisms, do misconceptions hinder students' concept mastery?

We can see that the TS-MC-DINA-H model can simultaneously identify which concepts a student has mastered and which misconceptions they hold using only single-tier multiple-choice responses. Methodologically, this approach offers several advantages over CTT, IRT, and previous CDM applications. First, it provides richer diagnostic information than a single total score. Figure 8 shows pairs of students with identical total scores who nonetheless have markedly different patterns of attribute mastery. For example, student 108 exhibits uniformly high mastery probabilities across A1–A6, whereas student 132 shows clear gaps on several attributes despite having the same raw score. Second, by combining option-level Q-matrix coding with hierarchy-aware constraints, TS-MC-DINA-H helps mitigate the parameter proliferation that typically arises when concept mastery is jointly diagnosed alongside a large set of misconception attributes, thereby improving identifiability and decreasing the risk of nonconvergence during estimation [42,44]. In this study, the design enabled stable classification of six concept attributes and 44 misconception attributes with acceptable global model fit, far exceeding the number of attributes typically modeled in prior CDM work. Third, compared with traditional two-tier diagnostic tests that require a second “reason” tier [90], the present approach delivers two-tierlike diagnostic power by using only single-tier multiple-choice responses, thereby reducing test length and administrative complexity while preserving actionable misconception diagnoses. For instance, item 10 uses options B, C, and D to capture three distinct misconceptions, while the keyed option reflects mastery of the required concepts.

Beyond its methodological innovation, TS-MC-DINA-H also offers practical benefits for educators. This dual-focus

diagnosis allows teachers to identify each student's specific misunderstandings and then target instructions to correct those errors. By extracting richer information from routine tests, the model helps bridge the gap between assessment and instruction, thereby enabling more focused remediation of misconceptions. For example, if many students in a class show impetuslike beliefs about motion or confusion between action-reaction pairs and balanced forces, teachers can plan focused activities that directly confront these ideas. At the same time, the learner-level mastery-misconception profiles can be summarized in individualized feedback reports that inform students about their strengths, remaining difficulties, and the misconceptions that are most likely to hinder their understanding. Such reports can also be linked to recommended learning resources that address the relevant concepts and misconceptions. Because the approach relies on conventional multiple-choice items and can be implemented with existing CDM software, it fits typical classroom constraints on testing time and administration while still yielding fine-grained, actionable diagnostic information. Moreover, the workflow is not domain bound; it can be extended to other physics topics (e.g., electricity and magnetism) and even to other disciplines where misconceptions are well cataloged, such as mathematics or chemistry.

Despite its contributions, this work has several limitations. First, the current workflow remains relatively labor intensive. Defining attributes, building and checking the option-level Q-matrix, and interpreting dual (mastery-misconception) outputs all require high demands on teachers' time and assessment literacy. Second, the difficulty distribution indicates that the instrument is somewhat easier than average for our sample and includes a small number of weak items. Although this is acceptable for our present goal of diagnosing common misconceptions in typical high school mechanics classrooms, future revisions should add more challenging items and refine or replace the weakest items to improve measurement precision for highly proficient students. Third, mapping each distractor to a specific misconception attribute relies on expert judgment, which introduces subjectivity and may affect diagnostic validity. When misconceptions overlap or item contexts are nuanced, the mapping can become ambiguous, thus further complicating interpretation. In addition, we did not implement a formal empirical Q-matrix validation procedure for either stage. Future work should conduct empirical Q-matrix validation and sensitivity analyses [41]. Finally, how to translate fine-grained diagnostic profiles into everyday classroom action requires further design research in real teaching contexts.

Given the rapid advances in artificial intelligence (AI), prior work has demonstrated that a teacher-GenAI collaborative workflow can lower the implementation threshold of cognitive diagnosis and offer feasible classroom support, thus suggesting a practical route for routine adoption [91].

This work demonstrates that AI provides a coherent pathway that aligns with the limitations above. With respect to workflow demands, AI can draft candidate attributes and option-level mappings [91] and automatically generate diagnostic items with plausible distractors [92], thereby reducing teachers' workload. With respect to the subjectivity problem, AI has the potential to draft option-misconception and attribute-misconception links, support multilabel tagging, refine them using response data, and flag ambiguous cases for human review (for related work, see Ref. [93]). With respect to classroom enactment, AI has the potential to translate fine-grained mastery-misconception diagnostic profiles into everyday classroom action by setting learning goals, designing classroom activities, providing learning resources, and proposing assessment plans [91]. In short, the next phase of this research will focus on an AI-supported pipeline that automates parts of model specification, standardizes misconception tagging,

and operationalizes profile-to-instruction handoffs within everyday physics lessons.

ACKNOWLEDGMENTS

This study was supported by the Beijing Education Science "14th Five-Year Plan" 2024 continuation grant (Grant No. CFDA24064). We thank Xin Li, Renaguli, Tuerxunayi, and Abudukaiyoumu for their assistance with data collection. The first author extends sincere thanks to Aerzuguli for continued support.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

-
- [1] Ministry of Education of the People's Republic of China, *General Senior Secondary Physics Curriculum Standards*, 2017 ed., revised 2018 (People's Education Press, Beijing, 2018) (in Chinese).
 - [2] NGSS Lead States, *Next Generation Science Standards: For States, By States* (The National Academies Press, Washington, DC, 2013).
 - [3] J. Clement, Students' preconceptions in introductory mechanics, *Am. J. Phys.* **50**, 66 (1982).
 - [4] R. F. Gunstone and R. T. White, Understanding of gravity, *Sci. Educ.* **65**, 291 (1981).
 - [5] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, *Am. J. Phys.* **53**, 1043 (1985).
 - [6] I. A. Halloun and D. Hestenes, Common sense concepts about motion, *Am. J. Phys.* **53**, 1056 (1985).
 - [7] M. McCloskey, A. Caramazza, and B. Green, Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects, *Science* **210**, 1139 (1980).
 - [8] M. McCloskey, A. Washburn, and L. Felch, Intuitive physics: The straight-down belief and its origin, *J. Exp. Psychol. Learn. Mem. Cognit.* **9**, 636 (1983).
 - [9] G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, Accommodation of a scientific conception: Toward a theory of conceptual change, *Sci. Educ.* **66**, 211 (1982).
 - [10] L. C. McDermott, Millikan lecture 1990: What we teach and what is learned—Closing the gap, *Am. J. Phys.* **59**, 301 (1991).
 - [11] National Research Council, *How People Learn: Brain, Mind, Experience, and School*, Expanded ed. (National Academies Press, Washington, DC, 2000).
 - [12] J. P. Leighton and M. J. Gierl, *Cognitive Diagnostic Assessment for Education: Theory and Applications* (Cambridge University Press, Cambridge, 2007).
 - [13] A. A. Rupp, J. Templin, and R. A. Henson, *Diagnostic Measurement: Theory, Methods, and Applications* (Guilford Press, New York, 2010).
 - [14] M. J. Gierl, J. P. Leighton, and S. M. Hunka, An NCME instructional module on exploring the logic of Tatsuoka's rule-space model for test development and analysis, *Educ. Meas.* **19**, 34 (2000).
 - [15] K. K. Tatsuoka, Rule space: An approach for dealing with misconceptions based on item response theory, *J. Educ. Meas.* **20**, 345 (1983).
 - [16] L. V. DiBello, L. A. Roussos, and W. F. Stout, Review of cognitively diagnostic assessment and a summary of psychometric models, in *Handbook of Statistics*, Vol. 26: Psychometrics, edited by C. R. Rao and S. Sinharay (Elsevier, Amsterdam, 2007), pp. 979–1030.
 - [17] A. A. Rupp and J. L. Templin, Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art, *Measurement* **6**, 219 (2008).
 - [18] J. de la Torre, DINA model and parameter estimation: A didactic, *J. Educ. Behav. Stat.* **34**, 115 (2009).
 - [19] J. L. Templin and R. A. Henson, Measurement of psychological disorders using cognitive diagnosis models, *Psychol. Methods* **11**, 287 (2006).
 - [20] J. de la Torre, The generalized DINA model framework, *Psychometrika* **76**, 179 (2011).
 - [21] T. Xu, X. Wu, S. Sun, and Q. Kong, Cognitive diagnostic analysis of students' mathematical competency based on the DINA model, *Psychologie* **60**, 3135 (2023).
 - [22] Y. Dong, X. Ma, C. Wang, and X. Gao, An optimal choice of cognitive diagnostic model for second language listening comprehension test, *Front. Psychol.* **12**, 608320 (2021).

- [23] B. C. Kuo, C. H. Chen, C. W. Yang, and M. M. C. Mok, Cognitive diagnostic models for tests with multiple-choice and constructed-response items, *Educ. Psychol.* **36**, 1115 (2016).
- [24] B. C. Kuo, C. H. Chen, and J. de la Torre, A cognitive diagnosis model for identifying coexisting skills and misconceptions, *Appl. Psychol. Meas.* **42**, 179 (2018).
- [25] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
- [26] R. Driver, A. Squires, P. Rushworth, and V. Wood-Robinson, *Making Sense of Secondary Science: Research into Children's Ideas* (Routledge, London, 1994).
- [27] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active-learning laboratory and lecture curricula, *Am. J. Phys.* **66**, 338 (1998).
- [28] L. Bao and E. F. Redish, Concentration analysis: A quantitative assessment of student states, *Am. J. Phys.* **69**, S45 (2001).
- [29] T. Martin-Blas, L. Seidel, and A. Serrano-Fernández, Enhancing Force Concept Inventory diagnostics to identify dominant misconceptions in first-year engineering physics, *Eur. J. Eng. Educ.* **35**, 597 (2010).
- [30] A. Savinainen and P. Scott, The Force Concept Inventory: A tool for monitoring student learning, *Phys. Educ.* **37**, 45 (2002).
- [31] A. Savinainen and P. Scott, Using the Force Concept Inventory to monitor student learning and to plan teaching, *Phys. Educ.* **37**, 53 (2002).
- [32] A. Savinainen and J. Viiri, The Force Concept Inventory as a measure of students' conceptual coherence, *Int. J. Sci. Math. Educ.* **6**, 719 (2008).
- [33] J. I. Yasuda and M. A. Taniguchi, Validating two questions in the Force Concept Inventory with subquestions, *Phys. Rev. ST Phys. Educ. Res.* **9**, 010113 (2013).
- [34] M. Planinić, L. Ivanjek, and A. Sušac, Rasch model based analysis of the Force Concept Inventory, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010103 (2010).
- [35] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, *Phys. Rev. ST Phys. Educ. Res.* **11**, 020134 (2015).
- [36] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010137 (2018).
- [37] G. W. Fulmer, Validating proposed learning progressions on force and motion using the Force Concept Inventory: Findings from Singapore secondary schools, *Int. J. Sci. Math. Educ.* **13**, 1235 (2015).
- [38] L. Bradshaw and J. Templin, Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions, *Psychometrika* **79**, 403 (2014).
- [39] M. A. Norris, Rethinking the Force Concept Inventory: Developing a cognitive diagnostic assessment to measure misconceptions in Newton's laws, Ph.D. thesis, Virginia Polytechnic Institute and State University, 2021.
- [40] K. Wancham, K. Tangdhanakanond, and S. Kanjanawasee, Sex and grade issues in influencing misconceptions about force and laws of motion: An application of cognitively diagnostic assessment, *Int. J. Instr.* **16**, 437 (2023), <https://e-iji.net/ats/index.php/pub/article/view/150>.
- [41] V. Le, J. M. Nissen, X. Tang, Y. Zhang, A. Mehrabi, J. W. Morphew, H.-H. Chang, and B. Van Dusen, Applying cognitive diagnostic models to mechanics concept inventories, *Phys. Rev. Phys. Educ. Res.* **21**, 010103 (2025).
- [42] G. Xu and S. Zhang, Identifiability of diagnostic classification models, *Psychometrika* **81**, 625 (2016).
- [43] J. de la Torre, A cognitive diagnostic model for cognitively based multiple-choice options, *Appl. Psychol. Meas.* **33**, 163 (2009).
- [44] M. Elbulok, A cognitively diagnostic modeling approach to diagnosing misconceptions and subskills, Doctoral dissertation, Columbia University, 2021.
- [45] S. Mi, J. Ye, Y. Li, and H. Bi, Development and validation of a conceptual survey instrument to evaluate senior high school students' understanding of electrostatics, *Phys. Rev. Phys. Educ. Res.* **19**, 010114 (2023).
- [46] J. X. Yao and C. S. He, *Improving Secondary School Physics Teaching Based on Learning Progressions: Kinematics and Force* (Beijing Normal University Press, Beijing, 2020) (in Chinese).
- [47] E. Kim and S. J. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, *Am. J. Phys.* **70**, 759 (2002).
- [48] M. Reiner, J. D. Slotta, M. T. H. Chi, and L. B. Resnick, Naive physics reasoning: A commitment to substance-based conceptions, *Cognit. Instr.* **18**, 1 (2000).
- [49] H. Ş. Kızılcık, M. Aygün, E. Şahin, N. Önder-Çelikkanlı, O. Türk, T. Taşkın, and B. Güneş, Possible misconceptions about solid friction, *Phys. Rev. Phys. Educ. Res.* **17**, 023107 (2021).
- [50] D. E. Brown, Students' concept of force: The importance of understanding Newton's third law, *Phys. Educ.* **24**, 353 (1989).
- [51] J. Klammer, An overview of techniques for identifying, acknowledging and overcoming alternate conceptions in physics education, ERIC Document No. ED423121, ERIC, 1998.
- [52] S. Gönen, A study on student teachers' misconceptions, and scientifically acceptable conceptions about mass and gravity, *J. Sci. Educ. Technol.* **17**, 70 (2008).
- [53] N. Sari, Murniati, and S. Ilyas, The implementation of problem-based learning modules to decrease misconception on Newton's law topic, *J. Phys. Conf. Ser.* **1460**, 012137 (2020).
- [54] C. Kruger, D. Palacio, and M. Summers, Surveys of English primary teachers' conceptions of force, energy, and materials, *Sci. Educ.* **76**, 339 (1992).
- [55] S. Bayraktar, Misconceptions of Turkish pre-service teachers about force and motion, *Int. J. Sci. Math. Educ.* **7**, 273 (2009).
- [56] A. Caramazza, M. McCloskey, and B. Green, Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects, *Cognition* **9**, 117 (1981).
- [57] A. E. Prescott and M. C. Mitchelmore, Teaching projectile motion to eliminate misconceptions, in *Proceedings of the 29th Conference of the International Group for the*

- Psychology of Mathematics Education* (PME, Melbourne, 2005), Vol. 4, pp. 97–104.
- [58] T. Neidorf, A. Arora, E. Erberber, Y. Tsokodayi, and T. Mai, *Student Misconceptions and Errors in Physics and Mathematics: Exploring Data from TIMSS and TIMSS Advanced* (Springer, Cham, 2020).
 - [59] A. Mutsvangwa, A study of student teachers' misconceptions on uniform circular motion, *J. Phys. Conf. Ser.* **1512**, 012029 (2020).
 - [60] J. de la Torre and N. Minchen, Cognitively diagnostic assessments and the cognitive diagnosis model framework, *Psicol. Educ.* **20**, 89 (2014).
 - [61] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, A review of multiple-choice item-writing guidelines for classroom assessment, *Appl. Meas. Educ.* **15**, 309 (2002).
 - [62] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/hrx4-gpd2> for the full original Chinese test and its English translation.
 - [63] L. Akbay and J. de la Torre, Estimation approaches in cognitive diagnosis modeling when attributes are hierarchically structured, *Psicothema* **32**, 122 (2020).
 - [64] W. Ma and J. de la Torre, GDINA: An R package for cognitive diagnosis modeling, *J. Stat. Software* **93**, 1 (2020).
 - [65] A. C. George, A. Robitzsch, T. Kiefer, J. Groß, and A. Ünlü, The R package CDM for cognitive diagnosis models, *J. Stat. Software* **74**, 1 (2016).
 - [66] D. Rizopoulos, LTM: An R package for latent variable modeling and item response theory analyses, *J. Stat. Software* **17**, 1 (2006).
 - [67] W. Revelle, *PSYCH (Version 2.5.6) [R package]* (Northwestern University, Evanston, IL, 2025).
 - [68] J. Chen, J. de la Torre, and Z. Zhang, Relative and absolute fit evaluation in cognitive diagnosis modeling, *J. Educ. Measure.* **50**, 123 (2013).
 - [69] M. Hansen, L. Cai, S. Monroe, and Z. Li, Limited-information goodness-of-fit testing of diagnostic classification item response models, *Br. J. Math. Stat. Psychol.* **69**, 225 (2016).
 - [70] W. Wang, L. Song, P. Chen, Y. Meng, and S. Ding, Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment, *J. Educ. Meas.* **52**, 457 (2015).
 - [71] R. L. Ebel and D. A. Frisbie, *Essentials of Educational Measurement*, 5th ed. (Prentice Hall, Englewood Cliffs, NJ, 1991).
 - [72] M. Young, B.-A. Cummings, and C. St-Onge, Ensuring the quality of multiple-choice exams administered to small cohorts: A cautionary tale, *Perspect. Med. Educ.* **6**, 21 (2017).
 - [73] F. B. Baker, *The Basics of Item Response Theory*, 2nd ed. (ERIC Clearinghouse on Assessment and Evaluation, College Park, MD, 2001).
 - [74] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd ed. (McGraw-Hill, New York, 1994).
 - [75] M. Tavakol and R. Dennick, Making sense of Cronbach's alpha, *Int. J. Med. Educ.* **2**, 53 (2011).
 - [76] W. A. Nicewander, Conditional reliability coefficients for test scores, *Psychol. Methods* **23**, 351 (2018).
 - [77] B. F. Green, R. D. Bock, L. G. Humphreys, R. L. Linn, and M. D. Reckase, Technical guidelines for assessing computerized adaptive tests, *J. Educ. Measure.* **21**, 347 (1984).
 - [78] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control* **19**, 716 (1974).
 - [79] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference*, 2nd ed. (Springer, New York, 2002).
 - [80] G. Schwarz, Estimating the dimension of a model, *Ann. Stat.* **6**, 461 (1978).
 - [81] Y. Liu, W. Tian, and T. Xin, An application of M2 statistic to evaluate the fit of cognitive diagnostic models, *J. Educ. Behav. Stat.* **41**, 3 (2016).
 - [82] W. Ma, Evaluating the fit of sequential G-DINA model using limited-information measures, *Appl. Psychol. Meas.* **44**, 167 (2020).
 - [83] R. C. MacCallum, M. W. Browne, and H. M. Sugawara, Power analysis and determination of sample size for covariance structure modeling, *Psychol. Methods* **1**, 130 (1996).
 - [84] A. A. diSessa, Knowledge in pieces, in *Constructivism in the Computer Age*, edited by G. Forman and P. B. Pufall (Lawrence Erlbaum Publishers, New Jersey, 1988).
 - [85] S. Vosniadou, The development of students' understanding of science, *Front. Educ.* **4**, 32 (2019).
 - [86] M. T. H. Chi, J. D. Slotta, and N. de Leeuw, From things to processes: A theory of conceptual change for learning science concepts, *Learn. Instr.* **4**, 27 (1994).
 - [87] G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, Accommodation of a scientific conception: Toward a theory of conceptual change, *Sci. Educ.* **66**(2), 211 (1982).
 - [88] J. Tuminaro and E. F. Redish, Elements of a cognitive model of physics problem solving: Epistemic games, *Phys. Rev. ST Phys. Educ. Res.* **3**, 020101 (2007).
 - [89] C. Wang, H.-H. Chang, and J. Douglas, Combining CAT with cognitive diagnosis: A weighted item selection approach, *Behav. Res. Methods Instrum. Comput.* **44**, 95 (2012).
 - [90] D. F. Treagust, Development and use of diagnostic tests to evaluate students' misconceptions in science, *Int. J. Sci. Educ.* **10**, 159 (1988).
 - [91] K. Jielili, C. W. Yang, and J. X. Yao, Research on the application of cognitive diagnosis based on generative artificial intelligence, *J. China Examinations* **9**, 64 (2025).
 - [92] S. K. Bitew, J. Deleu, C. Develder, and T. Demeester, Distractor generation for multiple-choice questions with predictive prompting and large language models, *arXiv:2307.16338*.
 - [93] S. Moore, R. Schmucker, T. Mitchell, and J. Stamper, Automated generation and tagging of knowledge components from multiple-choice questions, in *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Association for Computing Machinery, New York, 2024), pp. 122–133.