



Assessment of preservice physics teachers' knowledge of student understanding of force and motion

Lan Yang,^{1,2} Leheng Huang^{1,2}, Xianqiu Wu,^{1,2,§} Jianwen Xiong,^{1,2,‡}
Lei Bao,^{3,†} and Yang Xiao^{1,2,*}

¹Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter,
National Demonstration Center for Experimental Physics Education,

School of Physics, South China Normal University, Guangzhou, Guangdong 510006, China

²Key Laboratory of Atomic and Subatomic Structure and Quantum Control (Ministry of Education),
South China Normal University, Guangzhou, Guangdong 510006, China

³Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA



(Received 15 February 2024; accepted 3 May 2024; published 31 May 2024)

In physics education, a number of studies have developed assessments of teachers' knowledge of student understanding (KSU) of specific physics concepts with modified versions of existing concept inventories, in which teachers were asked to predict the popular incorrect answers from students. The results provide useful but indirect information to make inferences about teachers' knowledge of the misconceptions that students may be using in answering the questions. To improve the assessment of teachers' KSU, a new instrument is developed using a three-tier item design. The items were adapted from 17 questions from the Force Concept Inventory on force and motion. Each item was designed in three tiers, with tier 1 asking for teachers' own answers to the question to test their content knowledge, tier 2 asking for teachers' predictions of popular students' incorrect answers, and tier 3 asking for teachers' explanations of students' incorrect answers in an open-ended form. The three-tier design captures teachers' content knowledge, predictions, and explanations in a single item to allow explicit measures of teachers' own content knowledge and their KSU on students' misconceptions. The instrument was validated with preservice physics teachers, who were master-level graduate students in a normal university in China. The assessment results also suggest that the preservice teachers' KSU of force and motion was only moderately developed, and their content knowledge was uncorrelated with their KSU. In addition, a four-level progression scale of KSU was also developed, which categorized the preservice teachers into five proficiency groups.

DOI: [10.1103/PhysRevPhysEducRes.20.010148](https://doi.org/10.1103/PhysRevPhysEducRes.20.010148)

I. INTRODUCTION

There is widespread agreement that teachers' pedagogical content knowledge (PCK) has been considered a critical factor for effective teaching [1]. PCK was first introduced by Shulman and conceptualized as "the ways of representing and formulating the subject that make it comprehensible to others" [2]. In the field of science education, Magnusson *et al.* further conceptualized it as five elements: (a) orientations toward science teaching, (b) knowledge and beliefs about science curriculum, (c) knowledge and beliefs

about student understanding of specific science topics, (d) knowledge and beliefs about assessment in science, and (e) knowledge and beliefs about instructional strategies for teaching science [3]. Accordingly, teachers are expected to know and engage students' existing knowledge and representations, which could be further incorporated into effective instruction [4,5].

However, it was found that physics teachers may struggle to think through the perspective of students, such as understanding the origin and impact of students' misconceptions [6]. In a series of studies on teacher's knowledge about students, a number of instructional methods have been developed, which include reflection on teaching practice, group discussion, or analysis of student conceptual understanding and students' work [7–9]. Effective implementation of these instructional strategies requires a deep understanding of students' knowledge. Therefore, assessment of teachers' knowledge of student understanding (KSU) is essential for the effective implementation of teaching strategies.

Recent studies have used concept inventories to assess teachers' PCK. For example, high school teachers were

*Corresponding author: xiaoyang@m.scnu.edu.cn

†Corresponding author: bao.15@osu.edu

‡Corresponding author: jwxiong@scnu.edu.cn

§Corresponding author: xqwu@scnu.edu.cn

asked to predict the types and sources of students' preconceptions of electric circuits using an electric circuit diagnostic instrument (CDI) [10]. Teaching assistants' predictions of the most common incorrect answer choices of introductory physics students were also investigated using the Conceptual Survey of Electricity and Magnetism (CSEM) and the Force Concept Inventory (FCI) [9,11]. However, a simple comparison between students' responses and teachers' predictions of students' responses does not reveal if teachers have gained a deeper understanding of the sources of students' incorrect answers [12]. In a recent study, Kirschner's group adapted two items in the FCI to develop a paper-and-pencil test for assessing physics teachers' knowledge of student understanding (KSU) of science as facets of PCK [13]. The test items were in an open-ended format, which can probe richer information on teachers' understanding of students' misconceptions than closed-ended items. Inspired by this study, we adapted the FCI into a semi-open-ended format to probe if teachers can make correct inferences on students' misconceptions underlying common incorrect options. The goal of this study is to develop a test instrument based on the FCI for assessing physics teachers' professional KSU of force and motion (KSU-FM) in a more in-depth way and apply this tool to evaluate the KSU-FM of preservice physics teachers.

The progression of teachers' knowledge and capability regarding PCK has also been well studied. For example, Thompson investigated the development of a model-based inquiry learning progression for preservice teachers with a bachelor's degree [14–16]. Schneider and Plasman studied the teachers' developmental levels of PCK from preservice to lead teachers based on the five components of PCK proposed by Magnusson *et al.* [3,17]. On the assessment of levels of teachers' PCK, Schiering *et al.* used the scale anchoring procedure and identified four different proficiency levels based on data collected with teacher's PCK test from preservice physics teachers [18]. In this study, the FCI was adapted into a semiopen format test to evaluate teachers' KSU-FM, which was further analyzed to identify the developmental levels of teachers' KSU-FM.

II. LITERATURE REVIEW

A. Pedagogical content knowledge

Shulman first introduced the construct of PCK in his presidential address to the American Education Research Association (pp. 7–8) [2]. There is a general agreement that teachers' PCK is a crucial factor that affects teachers' teaching and student learning [19–21]. However, in the study of teachers' PCK, researchers hold different views on conceptualization and structure of PCK. A recent systematic literature review on PCK found that nearly 90% of the literature adopted a transformative perspective in conceptualizing PCK, in which PCK was classified into distinct categories of knowledge [22]. In the field of science

education, the commonly used PCK models include Magnusson *et al.*'s transformative model and its variants [3,23,24], which were further conceptualized into multiple PCK components. Among these components, the two most commonly agreed-upon and investigated PCK components include knowledge of students' understanding (KSU) and knowledge of instructional strategies and representations (KISR), which align well with Shulman's original proposal.

The main objective of this study is the assessment of one key component of PCK, i.e., teachers' KSU. In Magnusson *et al.*'s transformative model [3], this component of PCK includes teachers' knowledge about the science topic areas that students find difficult to learn, the prerequisite knowledge for learning-specific scientific knowledge, as well as variations in students' approaches to learning as they relate to the development of knowledge within specific topic areas. In a variant of Magnusson *et al.*'s transformative model, Park and Oliver suggested the PCK component of teachers' KSU [24], which was defined to include knowledge of students' conceptions of particular topics, learning difficulties, motivation, and diversity in ability, learning style, interest, developmental level, and need. KSU has also been emphasized by other theoretical frameworks. For example, the conceptual change approach and cognitive psychology literature have pointed to the importance of knowing and engaging students' existing knowledge and representations and of incorporating these existing ideas into teaching as critical principles for instruction [4,5,25]. In addition, Piaget emphasized "optimal mismatch" between student ideas and instructional design for cognitive conflict and desired assimilation and accommodation of knowledge and similar ideas have been put forward by other researchers [25,26].

In recent years, there has been an increasing number of studies on the assessment of teachers' KSU [9–12]. However, the assessment methods are mostly in an open-ended qualitative form, which is difficult to implement on a large scale. Therefore, in this research, we aim to conduct two areas of study: (i) develop a quantitative assessment tool that can effectively measure teachers' KSU and (ii) apply the assessment tool to conduct an in-depth evaluation of teachers' KSU.

B. Progression of teachers' PCK

Research has shown that teachers' pedagogical content knowledge also advances with their learning through professional training or reflections on their own teaching experiences. For example, a number of studies have focused on developing a model-based inquiry learning method for teachers to help them reflect on their own learning and classroom teaching practices [14–16]. Specific to the field of PCK, there is a widespread agreement that teachers' PCK plays a critical role in teaching and learning [1,27,28]. Teachers' professional development toward a

high level of PCK is essential for teachers to effectively plan, teach, and reflect on instruction [12,29,30].

With the recent attention to the development of teachers' PCK, researchers are attempting to model the progression of such PCK in two methods. One is a top-down process based on meta-analysis of theoretical and experiential studies to define a learning progression scale of teachers' PCK [17,31,32]. The other is a data-driven bottom-up approach based on the proficiency test data of teachers to quantitatively categorize levels of the progression in teachers' PCK [33]. For the former, Jin *et al.* [31] developed a learning progression-based scoring system to evaluate the extent to which teachers understand the knowledge essential for teaching the science topics. Besides, Schneider and Plasman [17] collected 91 research articles from 1986 to 2010, integrating and refining the five components of PCK (proposed by Magnusson *et al.* [3]) with different professional experiences including preservice, new, some experience, much experience, and leader. Then they applied a learning progression framework to model the development of teachers' PCK and identified the progression trajectories from novices to experts.

The data-driven approach does not assume any predetermined progression levels and draws from features of test data to develop a learning progression scale for teachers' PCK [18,34]. Schiering *et al.* [34] first presented a model of proficiency levels in preservice physics teachers' PCK in 2019 but lacked validity evidence. In 2022, Schiering *et al.* analyzed $N = 427$ observations of preservice physics teachers and identified four different proficiency levels in preservice physics teachers' PCK by utilizing the scale anchoring procedure [18]. In particular, participants' knowledge of students' understanding, instructional strategies, and curriculum can characterize both low and high proficiency, but knowledge of assessment is specific to higher proficiency. Their study also revealed how teacher education might promote transitions into higher proficiency levels and what PCK teachers at a specific level are likely to have.

However, the hypothetical model of PCK progression developed by Schneider and Plasman suggests that in addition to achieving gradual proficiency as a whole, teachers ought to develop their PCK in each individual component of PCK [17]. Therefore, an in-depth examination into the progression of teachers' PCK across the various components is required. In this research, we study a specific component of PCK, namely the teachers' KSU of physics concepts. The data-driven approach is used to model the progression levels of teachers' KSU by using data analysis to identify different proficiency levels through the scale anchoring procedure.

C. Data-driven approaches to investigate PCK

In existing research, four types of methods have been commonly used to measure teachers' PCK, which include

(i) written questionnaires and surveys, (ii) artifacts from teaching tasks (i.e., lesson planning), (iii) interviews, and (iv) lesson observations [22]. Among these methods, written questionnaires and surveys were the most convenient approach to assess teachers' PCK in large-scale studies, which will be discussed in more detail later.

For the method of artifacts, teaching and learning materials were taken from different phases of the teaching cycle, i.e., the preactivity phase of teaching, such as lesson plans (e.g., Bergqvist *et al.* [35]); the interactive phase of teaching, such as teaching videos (e.g., Chan and Yung [36]); and the postactivity phase of teaching, such as students' work and teachers' written reflections on the enacted lessons (e.g., Park and Oliver [24]).

The interview approach is often carried out with individual teachers [37–39] or focus group interviews [40,41]. Alonzo and Kim [40] recruited a teacher group to discuss class video clips on focus questions and then conducted interviews with the teachers. The study found that teachers' judgments were related to the quality of their discussions, and elaborated focus questions and interactions with colleagues may support novice teachers using their collective wisdom to engage in a situation-specific skill necessary for responsive teaching.

For lesson observations, teachers were observed live by the researcher(s). For example, through observing a lesson conducted by an experienced high-school physics teacher, Tay and Yeo identified eight pedagogical microactions that support the development of scientific models and modeling skills [42].

Finally, with the method of written questionnaires and surveys, participants provide written responses to a set of prompts (e.g., a pedagogical scenario) and/or questions and statements in a questionnaire or survey. Some items can be open ended, scenario based in the form of teaching vignettes. For example, Kind collected data via three topic-specific vignettes from 239 preservice science teachers to develop PCK rubrics [43]. Items can also be designed with a combination of different forms including true or false, multiple choice, matching, and short or long open response. For example, Sorge *et al.* developed a paper-and-pencil test to examine teachers' PCK regarding the force concept and Newton's laws [44].

Toward the assessment of the KSU component of PCK, researchers have developed open-ended questions to capture teachers' KSU, in which participants were asked to predict students' possible incorrect answers based on their understandings of students' misconceptions. For example, on the topic of multiplication of fractions, Isiksal and Cakiroglu [45] investigated teachers' knowledge of students' common misconceptions, the sources of such misconceptions, and the strategies to overcome the misconceptions. Similarly, Schmelzing *et al.* [46] developed a paper-and-pencil test in the setting of teaching biology, which asked participants to predict students' preconceptions and misconceptions.

When collecting data using written questionnaires and surveys, it is important that the validity of the original content questions is established. Such validity evaluates the extent to which a question can accurately probe students' common difficulties and/or misconceptions. It is reasonable to assume that teachers will be able to predict student's common difficulties only if the questions can accurately probe such difficulties. Therefore, many researchers use questions that have been previously validated as probes for assessing teachers' KSU (e.g., Zhou *et al.* [12]).

In physics education, many researchers utilize the existing conceptual surveys to study teachers' KSU, which are all previously developed and validated. For example, Lin [10] used an electric circuit diagnostic instrument to ask teachers to predict the preconceptions students may have about electric circuits and possible sources of the preconceptions. Maries and Singh [9] used the Force Concept Inventory (FCI) [47] to investigate teaching assistants' knowledge of students' difficulties in learning introductory physics. Karim *et al.* [11] adapted the Conceptual Survey of Electricity and Magnetism (CSEM) to evaluate teachers' knowledge of students' alternate conceptions.

The method of asking teachers to predict students' incorrect answers on a concept test has been used in many large-scale studies on evaluating teachers' KSU. However, this method does not explicitly probe teachers' understandings of students' misconceptions leading to their incorrect answers, which is considered to be the essential element of KSU [24]. Recent research also found that the teachers' PCK developed better when teachers had access to the learning progressions of students including their misconceptions [31,48]. Therefore, it is desirable to refine the prediction method to produce a more explicit measurement of teachers' understandings of students' misconceptions.

D. Research goals

As discussed in the literature review, methods of simply predicting students' incorrect answers cannot fully capture teachers' KSU, and it is important to develop measures that can directly probe teachers' understandings of the misconceptions behind students' incorrect answers. Therefore, in response to the limitations of the existing methods, this research aims to develop an assessment instrument that can explicitly probe physics teachers' knowledge of students' misconceptions on the topic of force and motion. The assessment instrument is then applied to examine preservice teachers' KSU of force and motion and determine a learning progression scale of the KSU based on the assessment data. This research is presented in three studies in this paper for clarity.

Study 1: Develop the assessment instrument on teachers' KSU of force and motion and evaluate its validity and reliability.

Study 2: Evaluate preservice teachers' KSU of force and motion.

Study 3: Develop a learning progression of teachers' KSU of force and motion.

By conducting these studies, this research seeks to provide insights into how teachers' KSU may call for the need for continued professional development. Through the analysis of the results, this research intends to offer valuable information that can inform efforts aimed at enhancing teacher education programs, improving classroom instruction, and supporting teachers in their efforts to meet students' learning needs.

III. METHODOLOGY

A. Design of the assessment instrument

The assessment instrument of this study, which is referred to as the test on Knowledge of Student Understanding on Force and Motion (KSU-FM), was developed by adapting 17 related items in the FCI. These 17 items were first identified by Neumann *et al.* [49] as an alternative instrument for validating the force and motion learning progression [50]. Each FCI item was extended into three questions in a three-tier structure, which are shown in the Supplemental Material [51]. The tier-1 question in an item is the original FCI question, which asks teachers to identify the correct answer to measure their content knowledge. The tier-2 question in the item asks teachers to choose the choice that they predict to be the most likely incorrect answer chosen by students. The tier-3 question is in an open-ended form which asks teachers to explain the students' misconception that leads to the incorrect answer selected in the tier-2 question.

Before the large-scale testing, a pilot study was conducted, which suggested that a time period of 45 min is an appropriate time frame for teachers to take the survey. The open-ended responses to the tier-3 questions were coded by two raters using a common coding scheme developed for each item. The coding scheme consists of misconceptions that were identified in the literature on the FCI [47]. In order to enhance the interrater agreement, the coding scheme includes multiple examples for a variety of explanations derived from teachers' responses in the pilot study. The initial percentage of agreement between the two raters was found to be approximately 83.8%. When an item has a discrepancy in coding, a faculty member will inspect the outcomes and discuss with both raters to reach a consensus for the final coding outcome.

B. Students' misconceptions on force and motion

The existing literature has documented a rich collection of students' misconceptions of force and motion [52,53]. In this research, we focus on the set of misconceptions targeted in the FCI [47], which are shown in Table VII in the Appendix along with the matching distractors of the test items.

C. Participants

The participants in this study were 86 first-year graduate students (preservice teachers) in the Master's program in physics education, who are trained to become physics teachers after completing their Master's program. All participants came from a normal university in China and had finished teacher education programs related to physics teaching and learning, including courses on physics, education theory, and pedagogical training. All of them had completed their physics courses, including mechanics, electricity and magnetism, optics, thermodynamics, and quantum physics in their undergraduate education. Besides, the participants all had a semester of physics teaching experiences in schools as part of their pedagogical training.

D. Data analysis

In this study, the three-tier question design was used to measure preservice teachers' content understanding (tier-1), their predictions (tier-2), and explanations (tier-3) of students' misconceptions. The three tiers of questions were scored dichotomously in separate ways. Tier-1 questions were scored with 1/0 for correct/incorrect answers to the content of the FCI questions, which will be called the "content score." Tier-2 and tier-3 questions were scored together, which will be called the "KSU score," where a score of 1 was assigned when the predicted students' incorrect answer (tier-2) was consistent with the identified students' misconceptions (tier-3). When the prediction and explanation were inconsistent or none of the misconceptions were identified, a score of 0 was assigned. Here, preservice teachers' content knowledge was evaluated using content scores based on the tier-1 questions. The results were also compared with students' KSU scores, which were obtained with tier-2 and tier-3 questions, to investigate the possible relations between content knowledge and KSU of force and motion.

1. Rasch analysis

Preservice teachers' KSU scores were further analyzed with Rasch analysis to establish the validity and reliability of the assessment tool and to provide evidence for determining a learning progression of preservice teachers' KSU of force and motion [54]. The Rasch model is a mathematical model developed by the Danish mathematician Georg Rasch around 1960, which has been widely applied in education assessment and physics education research [55]. The main objective of the model aims to have the estimates of item difficulty and person ability being mutually independent, which allows a common interval scale for direct comparison between item difficulty and person ability. Typically, the estimates of a person's ability and item difficulty are mapped on a single diagram, called a Wright map (see Figs. 1 and 3), for a visual representation of the distribution of person ability and item difficulty,

which can be further analyzed to determine whether an assessment instrument has a well-rounded capacity in assessing the targeted population. The Rasch dichotomous model was applied in the analysis for the KSU items. To perform Rasch modeling, the Winsteps 3.70 software was used.

2. Analysis of student misconceptions identified by the preservice teachers

The preservice teachers' ability to identify student misconceptions also varied across different misconceptions, which can provide useful information for determining content-specific indicators of KSU levels. There are a total of 23 misconceptions represented by the 17 questions of the KSU-FM test. The representations of these misconceptions are not uniform, i.e., some misconceptions have been represented in more questions than others. To measure how frequently a specific misconception is represented in the KSU-FM test, its frequency of representation is used, which is calculated with the ratio between the number of representations of the specific misconception and the total number of all misconception representations. Since each question may involve two or more misconceptions, the total number of misconception representations is larger than the number of questions of the KSU-FM test. Based on the FCI design [47], the 23 misconceptions are found to be represented 63 times among the 17 questions on force and motion.

Meanwhile, the preservice teachers' identifications of the misconceptions are also evaluated in terms of frequencies. The identification frequency of a specific misconception is calculated as the ratio between the number of identifications of the misconception by all participants and the total number of participants' responses to the KSU-FM test, which is 1462 (17 responses per person multiplied by 86 participants).

Since representations of the misconceptions in the KSU-FM test are not uniform and can impact teachers' identification frequencies, it is inappropriate to make direct comparisons of the absolute scales of teachers' identification frequencies due to the lack of a common representation baseline. To address this issue, the ratio between the identification frequency and the representation frequency (I/R ratio) is introduced to evaluate if a specific misconception is more or less likely identified by the teachers. The I/R ratio can be larger than 1 because the identification frequency is normalized based on teachers' responses and can be larger than the representation frequency which is normalized by the total representations in the KSU-FM test questions. If the I/R ratio is close to or larger than 1, it implies that the corresponding misconception is more likely identified by the teachers than other misconceptions. On the other hand, if the I/R ratio is much smaller than 1, it implies that the teachers often ignore or place less emphasis on the corresponding misconception.

3. Developing a learning progression of preservice teachers' KSU of force and motion

From the literature, there are two methods of identifying teacher developmental levels at KSU: one is based on theoretical and experiential summaries of teacher PCK [17,31,32], and the other is based on a proficiency model constructed from testing data to demonstrate the progression of teacher' PCK [18]. For KSU, there have not been established theories and empirical evidences that can determine item difficulties required for regression methods [56]. Toward the data-driven approaches, Schiering *et al.* used the scale anchoring procedure when defining the proficiency levels of teachers' PCK [18], which was proposed by Beaton and Allen in 1992 [57] and further extended by Mullis *et al.* [58,59]. In this study, we adopted the scale anchoring procedure used by Schiering *et al.* to model the development of teachers' KSU based on teachers' proficiency levels measured with KSU-FM. The method for determining the proficiency levels is shown in Table III, which is referred to as the Schiering's steps [18]. The details of the analytic procedures and data analysis outcomes are described in detail in Sec. IV C.

IV. RESULTS AND DISCUSSION

A. Study 1: Establish the validity and reliability of the assessment instrument

The KSU-FM was evaluated with a number of validity and reliability measures. First, content validity is often the most fundamental, which concerns the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose [60]. Content validity needs to be established in order to finalize the design of the assessment, whereas other aspects of validity, such as construct validity, can be analyzed with data from the implementation of the assessment. In this research, content validity was evaluated in two steps. First, the content knowledge of the test was considered to be valid since the questions were all adapted from the FCI. Second, three university physics faculty and two experienced physics teachers were consulted to provide feedback on the coding scheme used to code tier-3 questions for obtaining the KSU scores. In this step, the faculty and teachers were asked to compile a two-way checklist to include all the misconceptions and the corresponding distractors of the test items. Based on the checklist, the coding scheme was developed and discussed among experts to reach a final version.

1. Dimensionality

In this study, Rasch modeling was performed with the preservice teachers' KSU scores. As a result, Rasch analysis outcomes including person ability and item difficulties are all about KSU and have meaning identical to KSU ability and KSU difficulty, which may be used

interchangeably. To perform Rasch analysis, the test data need to satisfy requirements on unidimensionality. To examine the unidimensionality, a principal component analysis (PCA) of Rasch residuals and the standardized residual contrast plot were performed. This analysis is used to determine the size of the remaining variance after the Rasch dimension construct has been extracted. It was performed to examine whether the assessment data can be explained adequately by a single Rasch dimension. As recommended, the acceptable range of the eigenvalue of the first contrast of PCA, which is the first PCA component in the correlation matrix of the residuals, representing the largest secondary dimension, is less than 2.0 [61,62].

In this study, the PCA of the residual showed that the Rasch dimension explained 21.1% of the variance in the data, with its eigenvalue of 7.0. The first contrast (the largest secondary dimension) had an eigenvalue of 2.0 and accounted for 8.5% of the unexplained variance. The variance in the data explained by the Rasch measures was nearly 3 times more than the variance explained by the largest secondary dimension. Examination of the standardized residual plot did not reveal any items that were additionally clustered. Therefore, the results indicated that the data satisfied the assumption of unidimensionality of the Rasch model.

2. Item fit statistics and reliability

To explore how well the KSU score data fits the Rasch model, item fit statistics was performed, which generates two indicators of misfit including inlier-sensitive (infit) and outlier-sensitive (outfit) statistics. The infit mean square (MnSq) is sensitive to the pattern of responses to items targeted on the person's level, while the outfit mean square (MnSq) is more sensitive to responses to items with difficulty far from the person's level (i.e., outlier). For a good model fit, the Rasch model requires that both the infit and outfit will be close to 1.0. Values greater than 2.0 indicate significant differences from the model expectations whereas values less than 0.5 suggest that less information is being provided by the respondents due to less variation. For example, when low performers choose the correct answer to a very difficult item, the infit of the item can be close to 1.0, while the outfit can be greater than 2.0.

Table I shows the item parameters for each of the 17 items, which include the KSU score, item KSU difficulty, standard error of measurement, and fit statistics (both infit and outfit). Notice that the infit and outfit were also reported in standardized Z scores (Zstd), which report the significance of the (mis)fit. The Zstd converts the MnSq into an approximate t statistic that is more sensitive to sample size than MnSq values. Typically, the infit and outfit MnSq should be in the range of 0.60–1.40 and the Zstd should be in the range of -2.0 to 2.0 on Zstd [61].

The Rasch analysis of the test data shows that the infit MnSq is in the range of 0.88–1.14 while outfit MnSq is in

TABLE I. Item statistic of the KSU-FM from Rasch analysis. Item KSU scores were obtained based on the consistency between tier-2 and tier-3 questions, and item KSU difficulties were from Rasch analysis. The model S.E. is the standard error of the item difficulty estimated from Rasch analysis.

Item	Item KSU score	Item KSU difficulty	Model S.E.	Infit		Outfit	
				MnSq	Zstd	MnSq	Zstd
Q1	0.42	0.27	0.24	1.12	1.3	1.28	1.7
Q2	0.44	0.19	0.24	0.95	-0.5	0.88	-0.8
Q3	0.52	-0.29	0.25	0.93	-0.7	0.89	-0.8
Q4	0.65	-0.88	0.25	1.14	1.2	1.16	1.0
Q5	0.21	1.46	0.29	0.99	-0.0	1.20	0.7
Q6	0.57	-0.45	0.24	0.97	-0.3	0.93	-0.5
Q7	0.36	0.61	0.25	0.96	-0.3	0.87	-0.7
Q8	0.24	1.21	0.28	0.85	-1.1	0.67	-1.3
Q9	0.41	0.33	0.25	0.92	-0.8	0.85	-0.9
Q10	0.38	0.40	0.25	0.91	-0.9	0.98	-0.1
Q11	0.12	2.24	0.36	1.11	0.5	2.11	2.0
Q12	0.65	-1.04	0.26	1.00	0.1	0.99	0.0
Q13	0.70	-1.19	0.27	1.20	1.5	1.32	1.5
Q14	0.67	-1.03	0.26	0.88	-1.0	0.78	-1.2
Q15	0.50	-0.13	0.24	0.98	-0.2	1.04	0.3
Q16	0.67	-1.09	0.26	1.03	0.2	0.97	-0.1
Q17	0.60	-0.63	0.25	1.02	0.2	0.96	-0.2

the range of 0.67–2.11. The maximum value (2.11) is outside the desired range (0.60–1.40), indicating that there are items that may need additional inspection. Infit and outfit Zstd values range from -1.3 to 2.0, which is within the acceptable range of -2.0 to 2.0. An inspection of item statistics reveals that item 11 has an outfit MnSq value of 2.11 which is outside the desired range. Based on the KSU scores, this item 11 is the most difficult one in the test. Since the infit MnSq and Zstd values of item 11 are within the desired range, therefore, this item can be retained according to Yang *et al.* [63]. Overall, the fit statistics suggest that the test data have a satisfactory goodness of fit with the Rasch model.

In Rasch modeling, the reliability is evaluated with the person and item separation coefficients, which are calculated to be 1.47 and 3.44, respectively. These values correspond to Cronbach's alpha equivalent reliabilities of 0.68 and 0.92, which indicate satisfactory person reliability (> 0.65) and good item reliability (> 0.8) [64,65].

3. Person-item alignment

Ideally, the difficulties of test items are expected to align well with the subjects' abilities. To evaluate such alignment, a Wright map is often used. Figure 1 depicts the Wright map of the 17 items ranked according to person ability. It provides a graphical summary of the distribution of item difficulty and person ability that are expressed along the same interval logit scale. A Wright map is also useful for establishing construct representativeness and

determining the extent to which items align to the ability and for identifying locations of the Rasch scale that are in need of improvement. The center of the Wright map is a line representing a common logit scale. On the left-hand side of the scale are persons, sorted by their abilities (proficiency on the KSU-FM), with the most proficient teachers at the top. The right side of the scale shows the distribution of items ranked by item difficulty with the easiest item (Q13) at the bottom and the most difficult item (Q11) at the top.

Overall, the Wright map in this research indicates that most of the ability ranges (of teachers) are generally well covered, thus indicating the representativeness of test items. As shown in the graph, the estimates on both sides of the logit scale overlap substantially with both means close to 0 (ability mean = -0.12 and difficulty mean = 0). This suggests that all the items can be considered appropriately aligned to the teachers in the current sample such that the information contained in the items could allow for an accurate discrimination among teachers at different levels of their KSU ability [55]. In other words, the items were not too difficult or too easy for this group of teachers. However, the items appear to be mainly located in the middle range of the attribute, suggesting that the test provides the greatest amount of information for participants with medium or medium-to-high ability but may not discriminate well among people with very low ability on KSU. Overall, the Wright map indicates a reasonably good alignment, although additional easier items may also be included to better target teachers of lower ability levels.

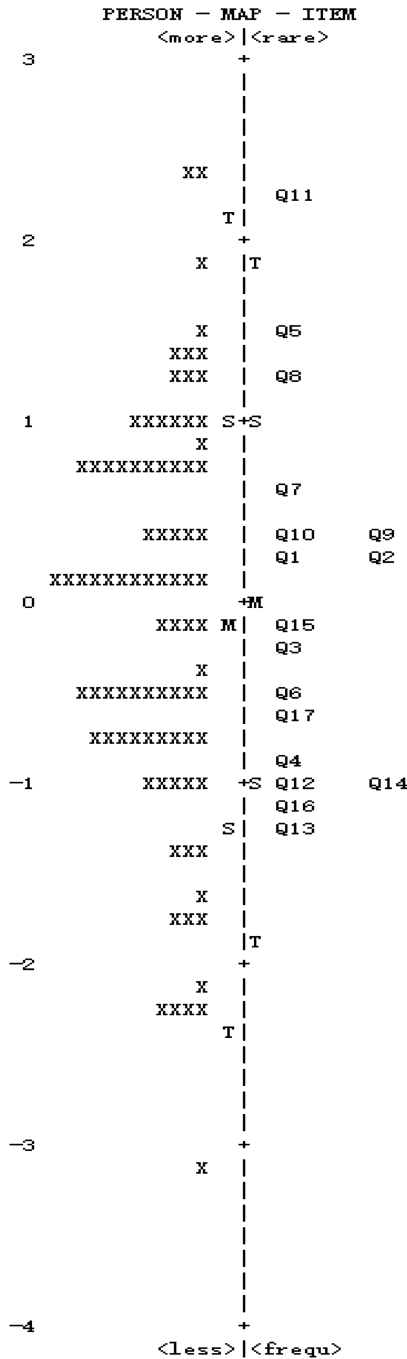


FIG. 1. Wright map of the KSU-FM test items. The vertical axis provides a scale of the estimated person ability and item difficulty. In the Rasch model, person ability and item difficulty are linear with the log of the probability of predicting a correct response, therefore, the scale of both person ability and item difficulty is in a “logit” unit. To the left of the axis, the individual students’ estimated person abilities are plotted with “x,” showing the distribution of students with different abilities. On the right side of the axis, the items are plotted based on their estimated item difficulties, which reveals the distribution of items across different difficulty levels. This map can conveniently show the distributions of both person abilities and item difficulties side by side, which allows a quick visual evaluation of the assessment features of the test items and the performance of students.

To summarize, based on experts’ evaluations and Rasch analysis, the KSU-FM survey appears to provide a valid coverage of the targeted content and can reliably discriminate teachers at different levels of KSU of force and motion.

B. Study 2: Preservice physics teachers’ KSU of force and motion

1. KSU score of KSU-FM and its correlation with content knowledge

The assessment of teachers’ KSU of force and motion was based on the tier-2 and tier-3 questions of KSU-FM. In these two questions, teachers were asked to predict students’ incorrect answers (in tier-2) and explain (in tier-3) their predictions by identifying the underlying misconceptions that students may have. The two questions were graded together to produce the KSU scores. Teachers received 1 point on each question for explaining their predictions with the appropriate student misconceptions. If a teacher made no predictions in the tier-2 question or did not explain the tier-2 prediction with the appropriate student misconceptions, the teacher will receive 0 points for the question set. The results revealed a medium-low average KSU score of 47.8% for the preservice physics teachers tested. From the evaluation point of view, the overall performance indicated that the teachers’ KSU on students’ misconceptions of force and motion was underdeveloped.

In studies on PCK, it is always important to examine if subjects’ content knowledge may influence their PCK. In KSU-FM, the tier-1 questions were the original FCI items, which provide a direct assessment of subjects’ content knowledge. For these preservice teachers, their average content score on tier-1 questions was 90.5%, which was near the ceiling, indicating a well-developed understanding of the correct physics concept of force and motion. The correlation between content scores and KSU scores was low and insignificant ($r = 0.143$, $p = 0.188$), suggesting little interactions between content knowledge and KSU. The results demonstrate that having content knowledge alone will not spontaneously lead to the development of KSU, and additional targeted training is needed.

2. Preservice teachers’ ability in identifying different student misconceptions

To examine teachers’ identifications of different misconceptions involved in the KSU-FM, the identification and representation frequencies, as well as the I/R ratio are plotted in Fig. 2 with the numerical values given in Table II. In Fig. 2, the horizontal axis lists the 23 misconceptions. For each misconception, two bars are used to show the identification and representation frequencies. The I/R ratio between the two frequencies is also plotted against a second vertical axis.

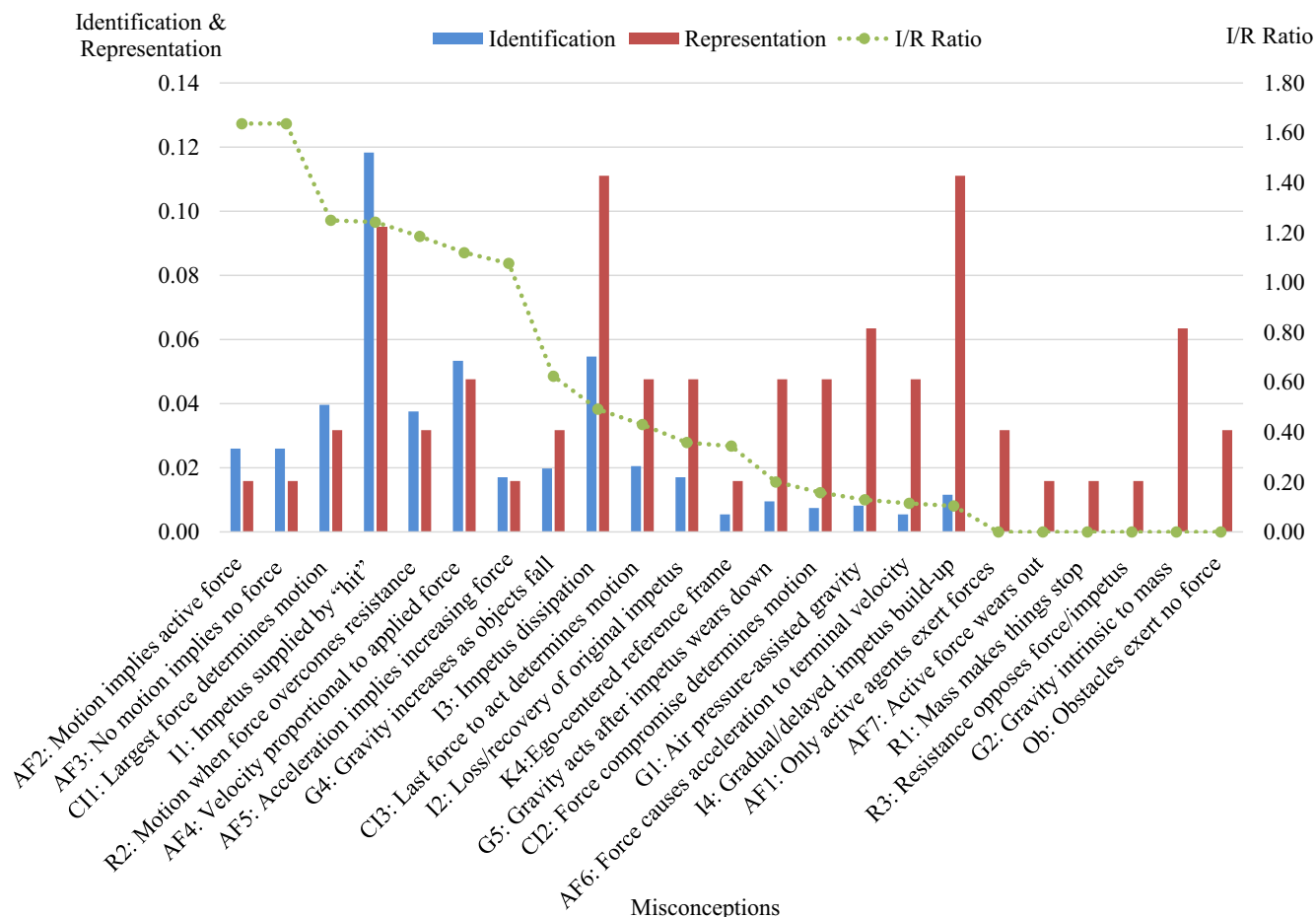


FIG. 2. Preservice physics teachers' identification of student misconceptions on force and motion. The identification frequency gives the fraction of teachers' total responses that explicitly mention a specific misconception. The representation frequency gives the fraction of the total answer choices of the KSU-FM test that target a specific misconception. The I/R ratio represents the ratio between the identification frequency and the representation frequency. Since the teachers' response frequency on a specific misconception is normalized by the total number of responses, it is possible for a response frequency to be larger than the representation frequency of the corresponding misconception, which leads to the ratio being larger than 1.

In Fig. 2 and Table II, the misconceptions are ordered from highest to lowest based on the I/R ratio. The results appear to reveal three groups of identification with distinctive gaps in I/R ratios between the adjacent groups. The I/R ratio values for dividing the different groups were chosen to be > 1.0 for the group of likely identified, 0.3 – 1.0 for the group of moderately identified, and 0.0 – 0.3 for the group of weakly identified. As shown in Table II, the mean values of the I/R ratios for the three groups are 1.31 , 0.45 , and 0.06 , respectively. One-way analysis of variance (ANOVA) with the Kruskal-Wallis nonparametric method showed significant differences in I/R ratios among the three groups ($\chi^2(2) = 19.1$, $p < 0.001$, $\epsilon^2 = 0.869$). Mann-Whitney tests show that the differences between adjacent groups are also significant ($p < 0.01$).

The likely identified group of misconceptions includes seven misconceptions (AF2, AF3, CI1, II, R2, AF4, and AF5). These misconceptions are mostly about the "active force" category, which is well documented in the literature

as being commonly observed among students learning introductory mechanics at high school and college levels [9,47].

The moderately identified group includes five misconceptions (G4, I3, CI3, I2, and K4), which represent a more diverse collection of students' thinking concerning the effects of motion from impetus, gravity, and reference frame. The weakly identified group includes the remaining 11 misconceptions (G5, CI2, G1, AF6, I4, AF1, AF7, R1, R3, G2, and Ob), which forms an even more diverse collection of different types of student thinking about the effects on motion from mass, gravity, competitions among forces, etc.

The results suggest that the preservice teachers were sensitive to the "active force" and "impetus" types of misconceptions and considered these misconceptions to be common among their students. In contrast, the preservice teachers' identifications of the remaining 11 misconceptions appeared to be inadequate. For example, the

TABLE II. Preservice teachers' identification of students' misconceptions of force and motion.

Misconceptions	Identification	Representation	<i>I/R</i> ratio	Identification groups	
AF2	Motion implies active force	0.026	0.016	1.64	Likely Identified <i>I/R</i> Mean = 1.31
AF3	No motion implies no force	0.026	0.016	1.64	
CI1	Largest force determines motion	0.040	0.032	1.25	
I1	Impetus supplied by "hit"	0.118	0.095	1.24	
R2	Motion when force overcomes resistance	0.038	0.032	1.19	
AF4	Velocity proportional to applied force	0.053	0.048	1.12	Moderately Identified <i>I/R</i> Mean = 0.45
AF5	Acceleration implies increasing force	0.017	0.016	1.08	
G4	Gravity increases as objects fall	0.020	0.032	0.62	
I3	Impetus dissipation	0.055	0.111	0.49	
CI3	Last force to act determines motion	0.021	0.048	0.43	
I2	Loss/recovery of original impetus	0.017	0.048	0.36	Weakly Identified <i>I/R</i> Mean = 0.06
K4	Ego-centered reference frame	0.005	0.016	0.34	
G5	Gravity acts after impetus wears down	0.010	0.048	0.20	
CI2	Force compromise determines motion	0.008	0.048	0.16	
G1	Air pressure-assisted gravity	0.008	0.063	0.13	
AF6	Force causes acceleration to terminal velocity	0.005	0.048	0.11	
I4	Gradual/delayed impetus build-up	0.012	0.111	0.10	
AF1	Only active agents exert forces	0.000	0.032	0.00	
AF7	Active force wears out	0.000	0.016	0.00	
R1	Mass makes things stop	0.000	0.016	0.00	
R3	Resistance opposes force/impetus	0.000	0.016	0.00	
G2	Gravity intrinsic to mass	0.000	0.063	0.00	
Ob	Obstacles exert no force	0.000	0.032	0.00	

misconception I4 is among the highly represented ones in the KSU-FM test, suggesting that this misconception was considered to be quite popular among students by the designers of the FCI. However, the teachers' identification of this misconception was very low with an *I/R* ratio of 0.10. The results suggest that these preservice teachers may lack knowledge about students' possible diverse range of understandings of force and motion, which is an important component of KSU for delivering effective instruction.

In addition, the variation in teachers' ability to identify different misconceptions provides valuable information that can be used for evaluating teachers' KSU and developing teacher training interventions. In the next section, the variation in teachers' identifications of the different misconceptions is used to develop a learning progression scale to evaluate teachers' levels of KSU of force and motion. The assessment goal is aimed to measure the extent to which teachers can identify the diverse range of misconceptions that students may have in thinking about force and motion.

C. Study 3: Developing a learning progression of teachers' KSU of force And motion

The basis for developing a learning progression is to categorize subjects into different groups that have distinctive performances in a progressive order. To do so, a technical procedure is to identify performance thresholds that can categorize subjects into different performance

levels. There is not a universally established method to determine such thresholds, which often need to be tailored to the specific context and conditions of a study, and the choice of methods will certainly impact the categorization of the groups. In research on PCK progression, the scale anchoring procedure implemented by Schiering *et al.* has been well established as a valid method to model a learning progression based on teachers' abilities from Rasch analysis [18]. This method will be used in this study to develop a learning progression of preservice teachers' KSU on force and motion based on their responses to the KSU-FM test (see Table III).

It is also noted that the learning progression developed in this study aims to reveal how preservice teachers' KSU on force motion may vary at different KSU levels, which can provide valuable information on whether certain misconceptions are more or less recognized among the preservice teachers at different KSU levels. This type of information can be further used in teacher training to develop more appropriately targeted instruction toward improving teachers' KSU. In addition, the result of this developed learning progression represents a cross-sectional outcome, which does not provide any developmental information and cannot be used to infer if the learners would actually move through these progression levels linearly as they learn. In general, learning should not be assumed to be a linear process that can be captured by a simple progression. Therefore, the learning progression developed in this study

TABLE III. The scale anchoring procedure implemented by Schiering *et al.* [18].

Schiering's method	Procedures
Step 1: Forming groups of participants	First, participants' Rasch abilities of KSU were transformed into a practicable scale from 300 to 700 points ($M = 514.08$, $SD = 75.89$). Three ranges (400–450, 500–550, and 600–650) were then defined along this scale, corresponding to groups 1, 2, and 3, respectively. These ranges were defined such that the resulting three groups each represent a sizable subsample but also are simultaneously small enough to ensure a homogeneous (within groups) and distinguishable (between groups) set of scaled abilities. The scale conversion was based on the method shown on page 556 in Linacre, J. M. (1993). <i>A User's Guide to BIGSTEPS: Rasch-Model Computer Program</i> . Mesa Press. or https://www.winsteps.com/winman/rescaling.htm
Step 2: Determining the group-wise proportion of correct answers	After defining different groups of participants, the groupwise percentages of participants who answered each item correctly were calculated. For example, in this study, item 14 was answered correctly by 25% of the participants in group 1, 71% of the participants in group 2, and 100% of the participants in group 3.
Step 3: Forming sets of items	Based on the groupwise percentages of correct answers, the items were grouped into different sets. The criteria used to group items according to their percentages of correct answers was adapted from Mullis and Fishbein [58,59]: <ul style="list-style-type: none"> • An item belongs to set 1 if at least 55% of participants of group 1 answered this item correctly. • An item belongs to set 2 if at least 55% of participants of group 2 and less than 50% of participants of group 1 answered this item correctly. • An item belongs to set 3 if at least 55% of participants of group 3 and less than 50% of participants of group 2 answered this item correctly. • An item belongs to set 3+ if less than 50% of participants of group 3 answered this item correctly.
Step 4: Computing the average item difficulties and characterizing levels	Finally, the average item difficulty of each of the four sets was computed, which defines four level thresholds used to categorize participants into five proficiency-level groups. To ensure that all level thresholds are able to distinguish the participants into distinctive proficiency level groups, it is tested that a typical person in proficiency level group i (i.e., a person with KSU ability equal to the level threshold i) would solve a typical item of level $i + 1$ (i.e., an item with difficulty equal to the level threshold $i + 1$) with a low success probability [56]. The mean KSU abilities of participants in different proficiency level groups were also tested for statistical significance.

was aimed to provide a cross-sectional outline of preservice teachers' KSU-FM knowledge and insights into their preparation, but it should not be used for indications of any developmental pathways.

Following Schiering's method shown in Table III, the first step is to group participants into different performance groups. To do so, the teachers' KSU abilities obtained with Rasch modeling were transformed into a scale from 300 to 700 ($M = 491$, $SE = 5.41$) [61], with which the proficiency levels were determined using Schiering's method (step 1). We defined three performance groups (groups 1, 2, and 3) whose participants' scaled abilities were within three ranges (400–450, 475–525, and 550–600), respectively. The range for each group and the gaps between groups were determined based on the requirements established in the existing studies [57,58]. These ranges were defined such that the resulting three groups each represent a subsample that is homogeneous within the group and distinguishable between groups.

The results of three performance groups with level thresholds were shown in Table IV, which were found to

be statistically significant [$\chi^2(2) = 71.7$, $p < 0.001$] with the Kruskal-Wallis test. Here the Kruskal-Wallis test was used, instead of one-way ANOVA, because the data were not normally distributed. In addition, the mean scaled abilities of different groups were evaluated for being statistically different. Since the data do not follow a normal distribution, the Mann-Whitney U test, which is a

TABLE IV. The mean values of person ability from teachers in three performance groups categorized based on the step 1 procedure of Schiering's method. The p values are from Mann-Whitney U test, which is a nonparametric alternative to t test.

Group	Range of scaled ability	N	Scaled ability		Difference and p values
			Mean	SD	
1	400–450	13	431.54	16.22	61.7, ($p_{1,2} < 0.001$)
2	475–525	31	493.26	14.27	65.4, ($p_{2,3} < 0.001$)
3	550–600	8	558.63	8.94	

nonparametric alternative to t test, was used. The results indicated that all adjacent group means of participants' scaled abilities differ significantly (see Table IV), ensuring that the chosen groups of participants represent different KSU levels.

It is noted that under this grouping method, only a fraction of the total participants (52 out of 86) were assigned to the three performance groups. The participants with scaled abilities in the gap between group ranges were not included. This approach ensures that the performance levels between the groups are distinctively different so that the data from the different groups can be used in further analysis in steps 2–4 of Schiering's method to determine the unique discriminative features of specific assessment items.

In step 2, using the three performance groups identified in step 1, for each KSU-FM item, the groupwise percentage of participants who answered the item correctly was calculated for each group. This process generated three percentages of correct answers for each of the 17 items in the KSU-FM test. For example, item 14 was answered correctly by 25% of the participants in group 1, 71% of the participants in group 2, and 100% of the participants in group 3.

The third step is to categorize items into different sets based on the groupwise percentages of correct answers from step 2. The criteria used to group items according to their percentages of correct answers were adapted from Mullis and Fishbein [58,59] and summarized below (see Table III):

- An item belongs to set 1 if at least 55% of participants of group 1 answered this item correctly.
- An item belongs to set 2 if at least 55% of participants of group 2 and less than 50% of participants of group 1 answered this item correctly.
- An item belongs to set 3 if at least 55% of participants of group 3 and less than 50% of participants of group 2 answered this item correctly.
- An item belongs to set 4 if less than 50% of participants of group 3 answered this item correctly.

This process generated a total of four sets of items with two items in set 1, seven items in set 2, six items in set 3, and two items in set 4 (see Table V).

As the final step, the mean values of item difficulties from Rasch analysis were calculated with items in each of

the four-item sets. These average item-set difficulties were then defined as the thresholds of four corresponding proficiency levels for categorizing the progression of teachers' KSU (see Table V). In the Rasch model, item difficulties and person abilities are on a common scale; therefore, thresholds of item sets can be used as proficiency levels to categorize performance groups. Based on the level thresholds determined previously, the total 86 students were sorted into five proficiency-level groups. For example, a person is assigned to proficiency level i if the person's KSU ability from Rasch analysis is between level thresholds i and $i + 1$. Specifically, a person is assigned to proficiency level 0 if the person's KSU ability is below the level 1 threshold. The results are plotted in Fig. 3.

Figure 3 is a modified form of the Wright map, in which the right side shows all KSU-FM items ordered by their Rasch item difficulties and colored according to the item sets, with specific items labeled at the horizontal axis. The thresholds for four proficiency levels are represented with four horizontal lines, which are calculated as the mean item-set difficulties. Meanwhile, the left side of Fig. 3 shows the distribution of teachers' KSU abilities estimated by the Rasch model. Based on their KSU abilities, the total 86 teachers were assigned into five proficiency level groups that are also shown with matching colors of the item sets except for level 0: proficiency level 0 (below threshold level 1, 15.1%), proficiency level 1 (between threshold levels 1 and 2, 18.6%), proficiency level 2 (between threshold levels 2 and 3, 36.1%), proficiency level 3 (between threshold levels 3 and 4, 29.1%), and proficiency level 4 (above threshold level 4, 2.3%).

To check if the KSU abilities of the different proficiency groups significantly differ from one another, the mean values of the abilities of the different proficiency groups are calculated and compared (see Table VI), which show significant differences between adjacent levels. In addition, applying the Rasch equation using person abilities and item difficulties, it was confirmed that a person with ability at threshold level i would have a probability of no more than 40% to successfully solve a typical item with a Rasch difficulty at threshold level $i + 1$ [56]. Therefore, based on the results from statistical tests and Rasch modeling, it is verified that all proficiency-level groups are statistically different from one another.

To gain deeper insights into how preservice physics teachers at different proficiency levels in KSU-FM may respond to typical test items, the common misconceptions identified in typical items in different levels were analyzed and detailed below, which can be used as indicators of behaviors for achieving certain proficiency levels.

At proficiency level 0, most (84.1%) of the teachers' responses were not able to correctly identify the related misconceptions. These responses either do not mention any misconceptions or call for misconceptions unrelated to the answer choices in tier-2 questions. For the remaining

TABLE V. Mean values of Rasch difficulties of items in four item sets determined following the methods in Mullis and Fishbein [58,59].

Item set	Number of items	Difficulty mean
1	2	-1.14
2	7	-0.64
3	6	0.50
4	2	1.85

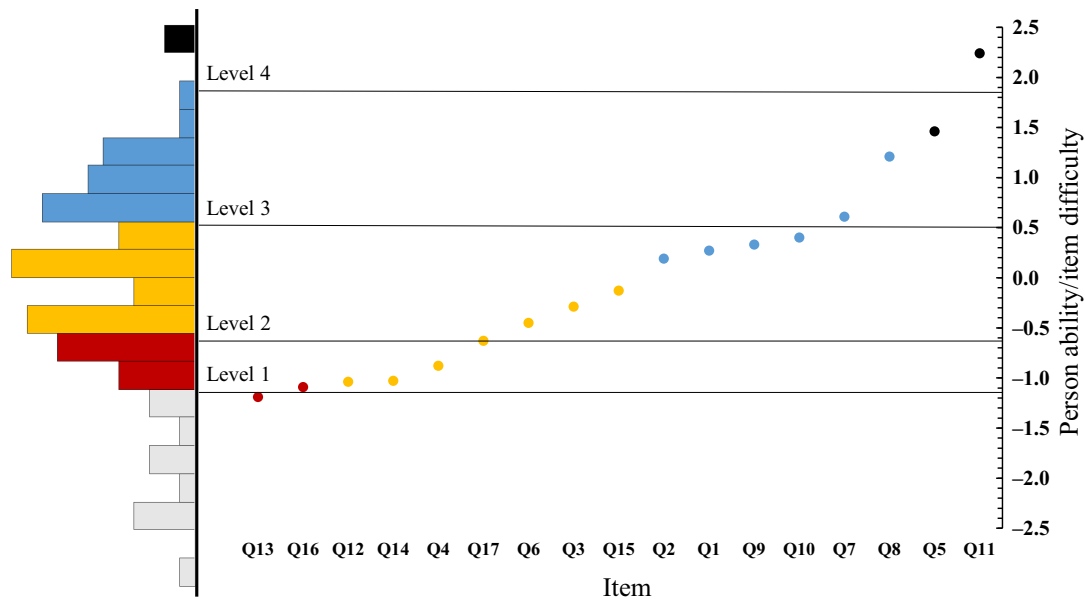


FIG. 3. Modified Wright map for KSU-FM test results. The dots represent items plotted based on their item difficulties and color coded to identify the four-item sets. The horizontal lines represent the thresholds of four corresponding proficiency levels, which are mean values of the item difficulties of different item sets. For example, the level 1 threshold is the mean value of difficulties of the items in item set 1, which includes Q13 and Q16 and is marked in red. The bar charts show the distributions of all preservice teachers ($N = 86$) at different KSU abilities, categorized into five proficiency level groups, from level 0 to level 4.

15.9% of the responses that did match the appropriate misconceptions, the referred misconceptions were sparsely scattered across all the different ones represented in the KSU-FM test without a systematic structure. The results suggest that preservice teachers at this level lacked a consistent view or a basic understanding of the possible student misconceptions of force and motion.

At proficiency level 1, there are still more than half (63.9%) of the teachers’ responses failing to correctly identify the related misconceptions. However, the improvement over teachers at level 0 is obvious. Among the correct responses, there are obvious patterns observed. These teachers can easily identify misconceptions R2 (motion when force overcomes resistance) and I1 (impetus supplied by “hit”) represented in the level 1 items (Q13 and Q16). Responses to higher-level items were often incorrect. The

results suggest that teachers at this level were starting to develop a basic understanding of the KSU of force and motion.

At proficiency level 2, teachers’ performance further improved with more than half (53.7%) of the teachers’ responses correctly identifying the appropriate misconceptions. As representative indicators, teachers were primarily able to identify misconceptions including AF4 (velocity proportional to the applied force), I1 (impetus supplied by “hit”), CI1 (largest force determines motion), and AF3 (no motion implies no force) in the typical level 2 items (Q14, Q4, Q17, and Q6). In addition, these teachers were doing very well on the level 1 items but had more difficulties on the level 3 items. The results suggest that teachers at this level had an improved understanding of the basic misconceptions of force and motion.

At proficiency level 3, a majority fraction (71.5%) of the teachers’ responses were correct. As representative indicators, teachers were primarily able to identify misconceptions including AF5 (acceleration implies increasing force), I3 (impetus dissipation), and CI3 (last force to act determines motion) in typical level 3 items (Q1, Q10, Q7). These teachers also performed very well on level 2 and level 1 items but had weaker performances on level 4 items. The results suggest that teachers at this level had more solidly developed understanding of an extended range of misconceptions about force and motion.

At proficiency level 4, most (88.2%) of the teachers’ responses were correct. However, with the participants in

TABLE VI. Preservice teachers’ KSU-FM Rasch abilities in different proficiency level groups.

Proficiency level groups	Number of teachers	KSU ability		p
		mean	SD	
Level 0	13	-1.97	0.498	$p_{0,1} < 0.001$
Level 1	14	-0.86	0.157	$p_{1,2} < 0.001$
Level 2	32	-0.08	0.317	$p_{2,3} < 0.001$
Level 3	25	1.01	0.312	$p_{3,4} = 0.021$
Level 4	2	2.36	0.000	

this study, only 2 out of the 86 achieved this level. Two level 4 items are also challenging. The preservice teachers at this level were able to identify the related misconceptions G5 (gravity acts after impetus wears down in Q5) and K4 (ego-centered reference frame in Q11). In contrast, the majority of the preservice teachers at other proficiency levels were not able to correctly identify these misconceptions and were categorized at or below proficiency level 3. The results suggest that items Q5 and Q11 can provide good discrimination to distinguish teachers for achieving well-developed KSU of force and motion.

In summary, the procedures to develop a progression scale of preservice teachers' proficiency levels KSU-FM and identify matched indicators in test items appear to be productive. In addition, the results also suggest that teachers' identification of students' misconceptions depends on the contextual scenarios presented in the items. For the same misconception, the difficulty of identification varies across different scenarios. However, in general, the misconception categories of "impetus" and "active force" were most commonly identified across different scenarios. The results suggest that for advancing assessment and training on KSU-FM, it is important to establish awareness among preservice teachers about the diverse nature of student misconceptions and place specific emphasis on less popular misconceptions that are poorly understood or ignored by many preservice teachers.

V. SUMMARY AND CONCLUSIONS

Being able to understand what misconceptions students have about physics is an important aspect of PCK, which empowers teachers to design and deliver effective pedagogical approaches that target these misconceptions and help students learn better. Past research has studied teachers' predictions of the incorrect answers that students may give [9,11], which provides evidence for implications of teachers' understanding of the student misconceptions underlying their incorrect answers. This study builds off the existing work to develop an explicit measure of teachers' KSU of force and motion with a three-tier question design that asks teachers to explain their predicted incorrect answers of students, which allows clear extractions of teachers' understandings of possible students' misconceptions.

Following the three-tier question design, an instrument for diagnosing teachers' KSU of force and motion (KSU-FM test) was developed and validated with Rasch analysis. Using the instrument, assessment data were collected with 86 graduate students (preservice teachers) in the physics education Master's program in a Chinese normal university. Descriptive statistics show that these students know the physics content very well with a mean content score of 90.5%. However, their KSU scores were much lower with a mean of 47.8%. The correlation between their content knowledge and KSU was also insignificant ($r = 0.143$,

$p = 0.188$). The results suggest that having a high level of content knowledge will not spontaneously develop KSU, and additional KSU-targeted training is needed.

Further analysis of the data shows that preservice teachers' ability to identify student misconceptions also varied across different misconceptions, with the "active force" and "impetus" categories of misconceptions being most commonly identified. In addition, the identifications of misconceptions were found to be context dependent, i.e., the identification of the same misconception varied with the different contextual scenarios in which the misconception was represented. These features suggest that the misconceptions and question contexts can both contribute to the difficulty levels of test items, which can be used to develop a progression scale to evaluate teachers' proficiency levels of KSU.

Using the scale anchoring procedure implemented by Schiering *et al.* [18], a progression of KSU-FM was developed based on teachers' abilities on KSU and item difficulties from Rasch analysis. The results identified four difficulty levels of test items and five proficiency levels of preservice teachers. Further analysis of responses from participants at different proficient levels and test items at different difficulties revealed a progression of development in KSU-FM and unique indicators in test items and misconceptions that are representative at different developmental levels. Such information can provide valuable utility for developing effective assessment and targeted teaching interventions on KSU-FM.

Certain limitations should be considered regarding the proposed progression of KSU-FM outlined in this study. First, the construction of threshold levels based on the scale anchoring procedure utilized an algorithm that, while proven to be effective and valid [18], was artificially generated and inherently contains some degree of arbitrariness that must be considered when interpreting the outcomes. For example, the first step of forming participant groups did not have a universally applicable optimal solution, which introduced a degree of arbitrariness in defining these groups and their range of abilities. Employing a different algorithm to define the groups will undoubtedly influence the results of progression levels, which is worth exploring in future research. In this study, we chose to use the method from the established work [18] to identify participant groups that offered sufficient sample sizes while remaining homogeneous and distinguishable in terms of KSU-FM, thereby forming the basis of a reliable scale anchoring procedure [56,57]. Second, it is important to note that the progression of KSU-FM proposed in this study has not yet been validated through alternative measures such as pre-post measures of development. In addition, it is crucial to acknowledge that learning can follow multiple pathways and should not be presumed to be a linear process easily captured by a simple progression. The learning progression developed in this study offers a

cross-sectional outline of preservice teachers' KSU-FM knowledge, but it should not be interpreted for indications of a linear developmental pathway. To enrich the current study, future research could involve designing teaching interventions based on KSU-FM progression and tracking the development of preservice teachers' KSU-FM to assess the extent to which it aligns with the assumptions of the proposed progression.

In summary, the new assessment design experimented in this study has been shown to be effective in probing preservice teachers' KSU of force and motion. Using the assessment, a number of useful results were obtained including a four-level progression scale. In addition, the assessment results suggest that the impetus and active force categories of misconceptions were most popularly identified across different scenarios by participants at different proficiency levels. In contrast, other less popular misconceptions seemed to have not been well understood by the preservice teachers tested. Therefore, it is suggested that teacher preparation and training programs should pay attention to establish the awareness among preservice

teachers about the broad collection of student misconceptions and target the ones that are less understood by preservice teachers.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Junpeng Zhang, Weiqing Zhan, and Tian Wang for their assistance with data collection. The authors would also like to thank Dr. Jane Jackson for her kindness in sharing the revised Table II for the Force Concept Inventory, which incorporates the taxonomy of naïve conceptions probed by the inventory. This work was supported by the National Social Science Foundation of China under Grant No. CHA200261. Any opinions expressed in this work are those of the authors and do not necessarily represent those of the funding agencies.

APPENDIX

Summary of misconceptions of force and motion.

TABLE VII. Misconceptions of force and motion probed by the KSU-FM test.

Category	Misconceptions on force and motion	KSU-FM item	
Kinematics	K4. Ego-centered reference frame	11A, B	
Impetus	I1. Impetus supplied by "hit"	2D; 4B, C; 8B, D, E; 12B, C; 15D, B; 16B, D, E	
	I2. Loss or recovery of original impetus	2C, E; 7A; 9A, D	
	I3. Impetus dissipation	3C; D; 5C, D; 9D; 10C, E; 11E; 12A, B, C; 15B	
	I4. Gradual or delayed impetus buildup	2D; 3B, D; 7D; 9E; 10D; 14C; 15E	
Active force	AF1. Only active agents exert forces	16A; 17E	
	AF2. Motion implies active force	4B, C; 9A; 12A; 15A	
	AF3. No motion implies no force	6E	
	AF4. Velocity proportional to applied force	7C; 8A; 12B; 14A	
	AF5. Acceleration implies increasing force	1B	
	AF6. Force causes acceleration to terminal velocity	1A	
	AF7. Active force wears out	8C, E	
Concatenation of influences	CI1. Largest force determines motion	13D, E; 17A, D	
	CI2. Force compromise determines motion	5A; 7C; 11C	
	CI3. Last force to act determines motion	2A; 7B; 9C	
Others	Ob. Obstacles exert no force	4A,B;6A	
	Friction	R1. Mass makes things stop	15A, B
		R2. Motion when force overcomes resistance	13A, B, D, E; 14B
		R3. Resistance opposes force or impetus	14B
	Gravity	G1. Air pressure-assisted gravity	1E; 4A; 6C, D; 17D
		G2. Gravity intrinsic to mass	1D; 4E; 6C; 12E
		G4. Gravity increases as objects fall	1B; 12B
		G5. Gravity acts after impetus wears down	5D, C, E; 11E; 12B

- [1] S. K. Abell, Research on science teacher knowledge, in *Handbook of Research on Science Education* (Routledge, London, 2007).
- [2] L. S. Shulman, Those who understand: Knowledge growth in teaching, *Educ. Res.* **15**, 4 (1986).
- [3] S. Magnusson, J. Krajcik, and H. Borko, Nature, sources, and development of pedagogical content knowledge for science teaching, in *Examining Pedagogical Content Knowledge: The Construct and its Implications for Science Education*, edited by J. Gess-Newsome and N. G. Lederman (Springer Netherlands, Dordrecht, 1999), pp. 95–132, [10.1007/0-306-47217-1_4](https://doi.org/10.1007/0-306-47217-1_4).
- [4] R. Duit and D. F. Treagust, How can conceptual change contribute to theory and practice in science education?, in *Second International Handbook of Science Education*, edited by B. J. Fraser, K. Tobin, and C. J. McRobbie (Springer Netherlands, Dordrecht, 2012), pp. 107–118, [10.1007/978-1-4020-9041-7_9](https://doi.org/10.1007/978-1-4020-9041-7_9).
- [5] D. Larkin, Misconceptions about ‘misconceptions’: Pre-service secondary science teachers’ views on the value and role of student ideas, *Sci. Educ.* **96**, 927 (2012).
- [6] L. Halim and S. Mohd. Mohd. Meerah, Science trainee teachers’ pedagogical Content Knowledge and its influence on physics teaching, *Res. Sci. Technol. Educ.* **20**, 215 (2002).
- [7] D. L. Hanuscin, Critical incidents in the development of pedagogical content knowledge for teaching the nature of science: A prospective elementary teacher’s journey, *J. Sci. Teach. Educ.* **24**, 933 (2013).
- [8] J. I. Heller, K. R. Daehler, N. Wong, M. Shinohara, and L. W. Miratrix, Differential effects of three professional development models on teacher knowledge and student achievement in elementary science, *J. Res. Sci. Teach.* **49**, 333 (2012).
- [9] A. Maries and C. Singh, Teaching assistants’ performance at identifying common introductory student difficulties in mechanics revealed by the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **12**, 010131 (2016).
- [10] J.-W. Lin, Do skilled elementary teachers hold scientific conceptions and can they accurately predict the type and source of students’ preconceptions of electric circuits?, *Int. J. Sci. Math. Educ.* **14**, 287 (2016).
- [11] N. I. Karim, A. Maries, and C. Singh, Exploring one aspect of pedagogical content knowledge of teaching assistants using the Conceptual Survey of Electricity and Magnetism, *Phys. Rev. Phys. Educ. Res.* **14**, 010117 (2018).
- [12] S. Zhou, Y. Wang, and C. Zhang, Pre-service science teachers’ PCK: Inconsistency of pre-service teachers’ predictions and student learning difficulties in Newton’s third law, *Eurasia J. Math. Sci. Technol. Educ.* **12**, 373 (2016).
- [13] S. Kirschner, A. Borowski, H. E. Fischer, J. Gess-Newsome, and C. von Aufschnaiter, Developing and evaluating a paper-and-pencil test to assess components of physics teachers’ pedagogical content knowledge, *Int. J. Sci. Educ.* **38**, 1343 (2016).
- [14] M. Braaten and M. Windschitl, Working toward a stronger conceptualization of scientific explanation for science education: Scientific explanations, *Sci. Educ.* **95**, 639 (2011).
- [15] J. Thompson, M. Braaten, and M. Windschitl, Learning progression as vision tools for advancing novice teachers’ pedagogical performance, in *Proceedings of the Learning Progressions in Science Conference, Iowa City, IA* (2009), pp. 1–25, https://www.researchgate.net/publication/272177786_Learning_Progression_as_Vision_Tools_for_Advancing_Novice_Teachers'_Pedagogical_Performance.
- [16] J. Thompson, M. Windschitl, and M. Braaten, Critical and contextual discourses: Explaining the development of ambitious practices across ‘learning-to-teach’ contexts, in *Proceedings of the Annual Conference of the National Association of Research in Science Teaching, Anaheim, CA* (2009).
- [17] R. M. Schneider and K. Plasman, Science teacher learning progressions: A review of science teachers’ pedagogical content knowledge development, *Rev. Educ. Res.* **81**, 530 (2011).
- [18] D. Schiering, S. Sorge, M. M. Keller, and K. Neumann, A proficiency model for pre-service physics teachers’ pedagogical content knowledge (PCK)—What constitutes high-level PCK?, *J. Res. Sci. Teach.* **60**, 136 (2022).
- [19] M. M. Keller, K. Neumann, and H. E. Fischer, The impact of physics teachers’ pedagogical content knowledge and motivation on students’ achievement and interest, *J. Res. Sci. Teach.* **54**, 586 (2017).
- [20] M. Kunter, U. Klusmann, J. Baumert, D. Richter, T. Voss, and A. Hachfeld, Professional competence of teachers: Effects on instructional quality and student development, *J. Educ. Psychol.* **105**, 805 (2013).
- [21] P. M. Sadler, G. Sonnert, H. P. Coyle, N. Cook-Smith, and J. L. Miller, The influence of teachers’ knowledge on student learning in middle school physical science classrooms, *Am. Educ. Res. J.* **50**, 1020 (2013).
- [22] K. K. H. Chan and A. Hume, Towards a consensus model: Literature review of how science teachers’ pedagogical content knowledge is investigated in empirical studies, in *Repositioning Pedagogical Content Knowledge in Teachers’ Knowledge for Teaching Science*, edited by A. Hume, R. Cooper, and A. Borowski (Springer, Singapore, 2019), pp. 3–76, [10.1007/978-981-13-5898-2_1](https://doi.org/10.1007/978-981-13-5898-2_1).
- [23] S. Park and Y.-C. Chen, Mapping out the integration of the components of pedagogical content knowledge (PCK): Examples from high school biology classrooms, *J. Res. Sci. Teach.* **49**, 922 (2012).
- [24] S. Park and J. S. Oliver, Revisiting the conceptualization of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals, *Res. Sci. Educ.* **38**, 261 (2008).
- [25] 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).
- [26] H. P. Ginsburg and S. Opper, *Piaget’s Theory of Intellectual Development*, 3rd ed. (Prentice-Hall, Inc, Englewood Cliffs, NJ, 1988), p. viii, 264.
- [27] J. Baumert, M. Kunter, W. Blum, M. Brunner, T. Voss, A. Jordan, U. Klusmann, S. Krauss, M. Neubrand, and Y.-M. Tsai, Teachers’ mathematical knowledge, cognitive activation in the classroom, and student progress, *Am. Educ. Res. J.* **47**, 133 (2010).

- [28] M. M. Keller, K. Neumann, and H. E. Fischer, The impact of physics teachers' pedagogical content knowledge and motivation on students' achievement and interest, *J. Res. Sci. Teach.* **54**, 586 (2017).
- [29] M. Krepf, W. Plöger, D. Scholl, and A. Seifert, Pedagogical content knowledge of experts and novices—what knowledge do they activate when analyzing science lessons?, *J. Res. Sci. Teach.* **55**, 44 (2018).
- [30] C. Kulgemeyer and J. Riese, From professional knowledge to professional performance: The impact of CK and PCK on teaching quality in explaining situations, *J. Res. Sci. Teach.* **55**, 1393 (2018).
- [31] H. Jin, H. Shin, M. E. Johnson, J. Kim, and C. W. Anderson, Developing learning progression-based teacher knowledge measures, *J. Res. Sci. Teach.* **52**, 1269 (2015).
- [32] G. J. Wiener, S. M. Schmeling, and M. Hopf, The technique of probing acceptance as a tool for teachers' professional development: A PCK study, *J. Res. Sci. Teach.* **55**, 849 (2018).
- [33] D. Schiering, S. Sorge, M. M. Keller, and K. Neumann, A proficiency model for pre-service physics teachers' pedagogical content knowledge (PCK)—What constitutes high-level PCK?, *J. Res. Sci. Teach.* **60**, 136 (2023).
- [34] D. Schiering, S. Sorge, S. Petersen, and K. Neumann, Konstruktion eines qualitativen Niveaumodells im fachdidaktischen Wissen von angehenden Physiklehrkräften, *Z. Didakt. Naturwiss.* **25**, 211 (2019).
- [35] A. Bergqvist, M. Drechsler, and S.-N. Chang, Rundgren, upper secondary teachers' knowledge for teaching chemical bonding models, *Int. J. Sci. Educ.* **38**, 298 (2016).
- [36] K. K. H. Chan and B. H. W. Yung, On-site pedagogical content knowledge development, *Int. J. Sci. Educ.* **37**, 1246 (2015).
- [37] A. C. Alonzo and J. Kim, Declarative and dynamic pedagogical content knowledge as elicited through two video-based interview methods, *J. Res. Sci. Teach.* **53**, 1259 (2016).
- [38] I. Henze, J. H. van Driel, and N. Verloop, Development of experienced science teachers' pedagogical content knowledge of models of the solar system and the universe, *Int. J. Sci. Educ.* **30**, 1321 (2008).
- [39] S. Walan, P. Nilsson, and B. M. Ewen, Why inquiry? Primary teachers' objectives in choosing inquiry- and context-based instructional strategies to stimulate students' science learning, *Res. Sci. Educ.* **47**, 1055 (2017).
- [40] A. C. Alonzo and J. Kim, Affordances of video-based professional development for supporting physics teachers' judgments about evidence of student thinking, *Teach. Teach. Educ.* **76**, 283 (2018).
- [41] D. F. Donnelly and A. Hume, Using collaborative technology to enhance pre-service teachers' pedagogical content knowledge in science, *Res. Sci. Technol. Educ.* **33**, 61 (2015).
- [42] S. L. Tay and J. Yeo, Analysis of a physics teacher's pedagogical 'micro-actions' that support 17-year-olds' learning of free body diagrams via a modelling approach, *Int. J. Sci. Educ.* **40**, 109 (2018).
- [43] V. Kind, Development of evidence-based, student-learning-oriented rubrics for pre-service science teachers' pedagogical content knowledge, *Int. J. Sci. Educ.* **41**, 911 (2019).
- [44] S. Sorge, J. Kröger, S. Petersen, and K. Neumann, Structure and development of pre-service physics teachers' professional knowledge, *Int. J. Sci. Educ.* **41**, 862 (2019).
- [45] M. Isiksal and E. Cakiroglu, The nature of prospective mathematics teachers' pedagogical content knowledge: The case of multiplication of fractions, *J. Math. Teach. Educ.* **14**, 213 (2011).
- [46] S. Schmelzing, J. H. van Driel, M. Jüttner, S. Brandenbusch, A. Sandmann, and B. J. Neuhaus, Development, evaluation, and validation of a paper-and-pencil test for measuring two components of biology teachers' pedagogical content knowledge concerning the 'cardiovascular system', *Int. J. Sci. Math. Educ.* **11**, 1369 (2013).
- [47] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [48] C. von Aufschnaiter and A. C. Alonzo, Foundations of formative assessment: Introducing a learning progression to guide preservice physics teachers' video-based interpretation of student thinking, *Appl. Meas. Educ.* **31**, 113 (2018).
- [49] I. Neumann, G. W. Fulmer, and L. L. Liang, Analyzing the FCI based on a force and motion learning progression, *Sci. Educ. Rev. Lett.* **8**, 8 (2013).
- [50] A. C. Alonzo and J. T. Steedle, Developing and accessing a force and motion learning progression, *Sci. Educ.* **93**, 389 (2009).
- [51] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.20.010148> for the test of teachers' knowledge of student understanding of force and motion.
- [52] I. Halloun and D. Hestenes, Common sense concepts about motion, *Am. J. Phys.* **53**, 1056 (1985).
- [53] L. Bao and E. F. Redish, Model analysis: Assessing the dynamics of student learning, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010103 (2006).
- [54] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 2nd ed. (Psychology Press, New York, 2007).
- [55] M. Planinic, W. J. Boone, A. Susac, and L. Ivanjek, Rasch analysis in physics education research: Why measurement matters, *Phys. Rev. Phys. Educ. Res.* **15**, 020111 (2019).
- [56] D. Woitkowski, Tracing physics content knowledge gains using content complexity levels, *Int. J. Sci. Educ.* **42**, 1585 (2020).
- [57] A. E. Beaton and N. L. Allen, Interpreting scales through scale anchoring, *J. Educ. Stat.* **17**, 191 (1992).
- [58] I. V. S. Mullis, K. E. Cotter, V. A. S. Centurino, B. G. Fishbein, and K. A. Reynolds, Using scale anchoring to interpret the TIMSS Advanced 2015 Achievement Scales, in *Methods and Procedures in TIMSS Advanced 2015* (2015). [Online]. Available at <https://timssandpirls.bc.edu/publications/timss/2015-a-methods/chapter-14.html>.
- [59] I. V. S. Mullis and B. Fishbein, Using scale anchoring to interpret the TIMSS 2019 Achievement Scales, in *Methods and Procedures: Timss 2019 Technical Report* (2019). [Online]. Available at https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch15-scale-anchoring.pdf.

- [60] S. N. Haynes, D. C. S. Richard, and E. S. Kubany, Content validity in psychological assessment: A functional approach to concepts and methods, *Psychol. Assess.* **7**, 238 (1995).
- [61] J. M. Linacre, A User's Guide to WINSTEPS® MINISTEP Rasch-Model Computer Programs Program Manual 3.80.0 (Winsteps.com, Beaverton, OR, 2012).
- [62] Y. Xiao, Kathleen Koenig, Jing Han, Q. Liu, J. Xiong, and L. Bao, Test equity in developing short version conceptual inventories: A case study on the conceptual survey of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **15**, 010122 (2019).
- [63] Y. Yang, P. He, and X. Liu, Validation of an instrument for measuring students' understanding of interdisciplinary science in grades 4-8 over multiple semesters: A Rasch measurement study, *Int. J. Sci. Math. Educ.* **16**, 639 (2018).
- [64] T. Bond, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd ed. (Routledge, New York, 2015), [10.4324/9781315814698](https://doi.org/10.4324/9781315814698).
- [65] J. F. Malec *et al.*, The Mayo High Performance Teamwork scale: Reliability and validity for evaluating key crew resource management skills, *Simul. Healthc.* **2**, 4 (2007).