Conceptual framework assessment of knowledge integration in student learning of measurement uncertainty

Chuting Lu,[1] Yating Liu,[1] Shaorui Xu,[2] Shaona Zhou,[1,†] Heather Mei,[3] Xiangqun Zhang,[3,4] Lan Yang,[1,3] and Lei Bao[3,*]

[1]*Guangdong 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 510006, China*
[2]*School of Electronics and Communication, Guangdong Mechanical & Electrical Polytechnic, Guangzhou, Guangdong 510550, People's Republic of China*
[3]*Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA*
[4]*Zhenjiang Experimental School, Zhenjiang, Jiangsu 212034, China*

(Received 31 October 2022; accepted 20 September 2023; published 16 October 2023)

In this study, a conceptual framework of measurement uncertainty was developed and used to guide the development of a multiple-choice concept test for the assessment of students' knowledge integration in learning measurement uncertainty. Based on assessment data and interview results, students were identified into three levels of knowledge integration including novice, intermediate, and expertlike. The reasoning pathways of students at different levels revealed a progression of reasoning from a rudimentary surface level to a deep understanding that can be mapped in the conceptual framework. This work demonstrates the possibility of identifying a quantitative categorization scheme to model knowledge integration as well as its utility in teaching and learning. Overall, the assessments and interviews revealed common and persistent difficulties in students' understanding of measurement uncertainty. In addition, students at different levels of knowledge integration demonstrate unique types of knowledge states that can be represented in the conceptual framework, making it a useful tool for analyzing different reasoning pathways and knowledge structures.

DOI: 10.1103/PhysRevPhysEducRes.19.020145

I. INTRODUCTION

A fundamental goal of physics education is for students to develop a deep understanding of essential scientific ideas [1,2]. Over the past few decades, investigating and improving students' conceptual understanding has become a fundamental goal in physics education [3–7]. However, many students lack a deep understanding of physics concepts after traditional instruction, leading to difficulty in applying their knowledge to solve novel problems. Traditional instruction often lends itself to rote memorization and its applications [8]. As such, students' obstacles can be difficult to overcome through traditional instruction, which does not change the ideas they developed from their everyday experiences and preconceived notions, and which are often incompatible with normative scientific ideas [9,10]. As a result, students may perform well on textbook problems with familiar contexts which they can solve using lower-level skills such as pattern matching of solutions and memorizing equations. However, they often fail to solve novel problems with unfamiliar contexts, which require students to have an integrated knowledge structure and deeper conceptual understanding [1,11–13]. The student learning behaviors exhibit the known characteristics from novice to expert knowledge structures, which can be modeled in terms of how their knowledge structures are constructed, activated, and linked [13–15].

The knowledge integration model typically distinguishes students into several developmental levels including novice, intermediate, and expert (or expertlike) [16–21]. Novice students often develop fragmented and disorganized knowledge structures, where knowledge is locally clustered with links connecting familiar contexts from personal experience and classroom learning. While solving problems, novices often match the surface features of problems with memorized processes and solutions. As a result, novices' applications of a concept are constrained to contexts similar to those encountered in classes or textbooks, leaving them unable to transfer their understanding to new situations [22–24]. Intermediate-level students have

*Corresponding author: bao.15@osu.edu
†Corresponding author: zhou.shaona@m.scnu.edu.cn

developed more connections in their knowledge structures but often fail to link the different knowledge components to the core principles (the central idea) of the concept and still partially rely on memorization at times. When solving problems, these students can demonstrate better understanding than the novices; however, the lack of understanding of the central idea limits their capability of transferring partial connections to novel scenarios. Therefore, they often fail to solve problems in unfamiliar contexts [16,25,26]. At the expertlike level, students' knowledge structures appear as integrated and hierarchically arranged networks with well-established links around a few core principles (the central idea). These students can achieve a deep understanding of the concept and use well-connected knowledge structures to solve problems in a wide range of settings [13,23].

To explicitly model students' knowledge structures and measure the levels of knowledge integration, the conceptual framework model was developed in previous studies [17–21]. A conceptual framework model usually consists of a central idea to serve as an anchor point and a range of related knowledge components such as contextual features and intermediate reasoning and processes. Contextual features in the conceptual framework can activate students' ideas and links. Each unique pathway connecting different contexts and conceptual elements of the framework can illustrate and model student learning, which helps researchers and teachers visualize how students structure certain concepts. Meanwhile, the differences in knowledge structures between novices and experts can be illustrated by the usage of central ideas and the connections among different knowledge components. The experts use the central idea as an anchor point to link related knowledge components, which extends to an integrated and hierarchical knowledge structure. The expert approach links a wide variety of contexts to the central idea, which can meaningfully and efficiently solve problems in different situations. Alternatively, the novices often bypass the central idea and directly link equations or algorithms to the surface features. Thus, novices can solve problems with familiar contexts but often fail in novel situations.

As shown from the previous studies [17–21], the conceptual framework model can be used as an operational guide to develop assessment instruments, which can probe different pathways within students' knowledge structures to reveal their levels of knowledge integration. The assessment results can then help transform classroom instruction to emphasize specific connections so that students can gain a deeper understanding and build integrated knowledge structures. The conceptual framework model has been developed and applied to several physical topics, such as light interference [17], force and motion [18], momentum [19], Newton's third law [20], and mechanical wave propagation [21]. In this research, the conceptual framework model is developed to examine students' understanding of measurement uncertainty in lab experiments, which is a fundamental learning goal in physics lab courses [27].

Physics is an experimental-based discipline, and physics knowledge taught to students has a strong experimental basis, which requires a good understanding of the uncertainties in measurements. The American Association of Physics Teachers [28] has prominently described one of the goals of physics teaching as understanding the nature of scientific measurement and uncertainty. However, a large number of students only display a rudimentary understanding of measurement uncertainty even after the completion of traditional laboratory courses [29–32]. Traditional laboratory courses often provide students with a laboratory manual to verify various laws, measure specific variables, or learn to use and become familiar with an experimental apparatus with the help of instructors. However, these practices of measurement and the intended learning of the concept of uncertainty are often difficult for students to grasp [33]. As a result, the routine training tasks in traditional laboratory courses reinforce the students' belief in the existence of a true value; the uncertainties in measurements are seen as errors but not as the intrinsic property of all measurements [34]. In other words, students commonly believe that, in principle, a perfect measurement without any uncertainty can be made [35]. In addition, students typically have no understanding of the need to make repeated measurements and often hold a general notion that repeated measurements bring about a better result, without understanding what the "better result" actually means [36]. In operation, students always treat the arithmetic mean as the final result of a dataset, which is all that matters when comparing two datasets [35]. These misunderstandings often make students unable to distinguish between uncertainties of random and systematic origins and fail to identify different sources of uncertainty in a measurement [33,36,37]. All these difficulties suggest that it is a challenge for students to achieve deep understanding and develop an integrated knowledge structure in learning measurement uncertainty.

Although several studies have focused on students' understanding of measurement uncertainty [30,33,37–43], very few researchers have measured students' conceptual understanding based on knowledge integration. In this research, a conceptual framework of experimental uncertainty is developed and applied to design an instrument for assessing students' understanding of measurement uncertainty. This leads to two areas of research that are conducted in this study:

(1) Develop a conceptual framework model of measurement uncertainty and use the conceptual framework to analyze student difficulties through the perspective of knowledge integration.
(2) Apply the conceptual framework to develop a multiple-choice assessment involving typical and

atypical contexts to evaluate students' levels of knowledge integration and deep understanding.

II. METHOD AND DESIGN

A. Student difficulties in learning measurement uncertainty

Over the last several decades, many researchers in physics education have documented rich information on students' common difficulties with measurement uncertainty [29–33,37–43] and evaluated the effectiveness of new instruction [44–46]. These studies reveal that students have a very limited understanding about the nature of uncertainty, which leads to different views on how uncertainty occurs and may be dealt with.

As shown by numerous studies, many students construct naive views of measurement uncertainty [33,37,40,43–45,47]. Some students will make arbitrary judgments about the measurement results and ignore the estimation of uncertainty [33]. These students lack the basic understanding of the need and process to determine measurement uncertainty. If students obtain different measurement outcomes, they often attribute the differences to human error, rather than something inherent in the measurement itself. Thus, most students believe that making a perfect measurement is possible when more advanced instruments are available and used by experts [33,34,40,48]. These naive views lead students to rarely conduct repeated measurements spontaneously unless there is something wrong with the first measurement or they get a value significantly different from their expectation [48]. A small number of these students would carry out multiple trials per their instructors' requirements. These students take repeated measurements as an operational routine [47] or believe that practice will make measurement perfect [49]. Therefore, these students tend to regard the value of a properly conducted first measurement as the final result value, where they believe that a single measurement can be perfect; or they can choose a recurring value being the final result, where they believe that getting the same value twice or more indicates well-conducted measurements [49]. In addition, students often compare datasets using a value-by-value comparison based on the closeness of each value, without attending to the uncertainties of the measured values [47]. These naive understandings indicate the lack of even a basic understanding of the origins of different types of uncertainties, as well as how uncertainties may be processed to yield meaningful measurement outcomes.

For more advanced students, they have developed a better understanding of uncertainty by establishing more connections among their knowledge components. For example, these students are aware of the influence of uncertainty on measurement results and understand some basic processes to work with uncertainty. However, these students still have difficulties in interpreting and analyzing measurement outcomes with multiple sources of uncertainties. That is these students do well in using equations or rules to calculate the results, but they lack the understanding and reasoning to explain how uncertainties of different origins contribute to measurement outcomes in complex settings [33,41]. For example, many students believe it is necessary to make repeated measurements and should always use the arithmetic mean to obtain the final result from a dataset [48]. However, they may calculate the average of a dataset without considering the process for rejection of anomalous values [40,41,49]. Besides, many students lack a basic understanding of sample size and its impact on measurement uncertainty. They typically believe that three is a default "good" number for measurements even though some lab materials may have stated that three measurements are insufficient. These students tend to compare different datasets based on only their mean values without considering standard deviations or errors since they believe that the average is all that matters [37,40,48]. In addition, these students often have difficulty identifying the primary source of uncertainty and distinguishing between random and systematic uncertainties [33,37,40]. Meanwhile, these students also demonstrate inconsistent understandings of uncertainty across different physics contexts and with different types of uncertainties [42,43], which suggests that the knowledge structures of these students are fragmented with partially connected local links such that different contexts may activate different local links that can be inconsistent viewed from an expert.

From reviewing the literature and current curricula on measurement uncertainty, it appears that traditional instruction often teaches measurement uncertainty in a narrow domain of context with emphasis on random uncertainty and mathematical calculations to address random uncertainties. As a result, systematic uncertainty is often treated as an unknown constant in the calculation, and students often focus on analyzing data using the averaging method with repeated measurements [44,45]. In addition, certain rules of thumb and *ad hoc* prescriptions, such as always making three measurements and taking an average, are often introduced in instruction without developing a good understanding of the underlying mechanisms, which usually leads students to merely memorize the rules and apply them in solving problems. In response to the students' learning difficulties and the limitations in the existing curricula, it is important to develop an integrated conceptual framework of measurement uncertainty, which can be used as a tool to represent and analyze different types of students' understandings and aid assessment and instruction.

B. The conceptual framework of measurement uncertainty

It appears that the current textbooks often focus on data processing to address measurement uncertainty [50] but

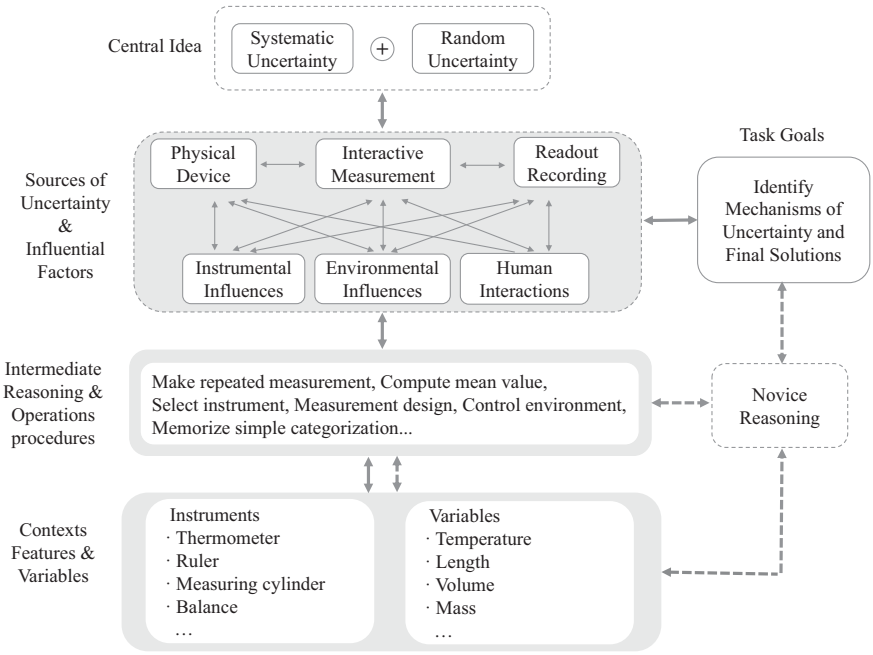

FIG. 1. Conceptual framework for the measurement uncertainty. Solid arrows represent experts' conceptual pathways while dashed-line arrows represent novices' reasoning.

have less emphasis on the conceptual understanding of the mechanisms that lead to the different types of measurement uncertainties. Meanwhile, very few studies have developed assessment instruments that target student understanding of the fundamental mechanisms of measurement uncertainty. In the existing literature, students' understandings of uncertainty were analyzed and interpreted based on the framework of data analysis and processing [37,39,41,48]. However, the empirical evidence from the existing literature has also demonstrated that students lack a basic understanding of the nature of uncertainty, even though they are able to solve computational questions defined in a narrow domain of context [33,41]. To directly address the root of students' learning difficulties, the conceptual framework model can be applied, which emphasizes the fundamental mechanisms of a concept defined as the central idea. In this section, a conceptual framework is developed to present the central idea and mechanisms of uncertainty, which is then used to model students' difficulties in terms of their understanding of the mechanistic nature of measurement uncertainty.

The first step to develop the conceptual framework of a concept is to identify the central idea based on experts' normative views [17–21]. As for measurement uncertainty, the normative views consider that measurement is an interaction between instruments and the entities being measured. A measurement setting involves humans, the environment, and instruments, which would interact with each other and lead to a range of measurement uncertainties that can be categorized into two types: systematic uncertainty and random uncertainty. The systematic uncertainty is a consistent shift from the presumed actual value in measurement outcomes, while the random uncertainty describes the random inconsistency in measurement outcomes. The possible mechanism leading to systematic and random uncertainties in a specific setting is context dependent and needs to be analyzed case by case. In general, both types of uncertainties are expressed in the processes of making and using an instrument. However, in most examples and practices, systematic uncertainty is often associated with certain properties of an instrument. Such instrumental properties are often fixed or slow changing, and therefore, are more likely to create a consistent shift in measurement outcomes than human interactions, which often lead to random uncertainties.

Following the definition of uncertainty, the mechanistic origin of measurement uncertainty is defined as the central idea, which consists of system interference and random processes. In the teaching and learning of measurement uncertainty, the central idea is closely linked to the sources of uncertainty and the methods of data analysis used to deal with the different kinds of uncertainties arising from measurements. Anchored with the central idea, a conceptual framework of measurement uncertainty is developed and shown in Fig. 1, which includes a range of additional knowledge components including contextual features, intermediate reasoning and operations, and different pathways of student reasoning in learning measurement uncertainty.

As represented in the conceptual framework, the knowledge elements (e.g., context features, intermediate reasoning, and sources of uncertainty) are organized hierarchically and link to the central idea. The top layer

component is the central idea, which builds on the core understanding of the mechanisms of uncertainty.

The second layer represents a concrete expression of the central idea in terms of an extended network of possible interactive relations in a measurement setting, which consists of different sources of uncertainty (device, interactive measurement, and readout process) and the influence factors (instrumental influences, environmental influences, and human interactions). Experts are usually able to identify the correct interactive pattern between sources of uncertainty and influencing factors, which helps them apply the central idea explicitly and intuitively to all related problems. In contrast, novice students often have a very limited understanding of this interactive relational network.

The third layer contains the intermediate reasoning processes and operational procedures, including mathematical, logical, and manipulative processing. These reasoning and mathematical manipulations provide operational rules and procedures to address measurement uncertainty, which are expected to be linked to the understanding of the origins of uncertainty. However, among novice students, these rules and procedures are often disconnected from deeper-level mechanistic understandings and applied as memorized procedures in problem solving. The common manipulations, such as making repeated measurements to calculate a mean value, selecting an instrument, or controlling the environment, are often explicitly taught in classes and can be learned by students as local links between contexts and operations based on memorization or through some of the naive type of intermediate processes. Besides, students tend to develop simple one-on-one rules for categorizing uncertainty, such as considering that instruments have only systematic uncertainty and random uncertainty is only with human observations.

The bottom layer consists of contextual features and variables, which are usually design features of problems and can be modified to create different questions and task settings. This layer represents the most concrete elements of the conceptual framework, which include surface details of context features such as the specific objects being measured and the measurement variables including temperature, length, volume, or mass, and the instruments involved, e.g., thermometer, ruler, measuring cylinder, and balance scale.

The conceptual framework integrates these layers and task goals to visually represent students' understandings in terms of reasoning pathways, which are shown as different links connecting the knowledge components at different layers (see the arrows shown in Fig. 1). Different contextual features, operations, sources, and influence factors are connected by arrows to represent the possible reasoning pathways of students. Solid arrows represent experts' reasoning pathways, while the dashed-line arrows represent the conceptual pathways of novices. Novices' reasoning is often based on surface-level context features of problems. They solve the problems by matching these features to operations encountered through textbooks and classes, without a deeper understanding of the central idea of uncertainty. Therefore, these students often have difficulties in answering questions with unfamiliar contexts or ones that require the understanding of the central idea. Experts, on the other hand, have developed an integrated knowledge structure that connects contextual features, operations, and the interactive pattern between sources and influence factors to the central idea. This knowledge structure enables them to reason from any given point of contextual features to reach the central idea, which allows them to develop problem-solving strategies with a range of flexible and comprehensive networks of reasoning pathways to successfully solve problems in both familiar and novel contexts.

C. Levels of knowledge integration within student knowledge structures

After developing the conceptual framework, students' difficulties and misconceptions documented in the existing literature can be interpreted and represented with different reasoning pathways of specific learning states within the framework to analyze students' level of knowledge integration. According to previous studies [17–21], asking students to solve problems with typical and atypical contexts in the assessments can distinguish between students at different levels of knowledge integration. Based on the conceptual framework of measurement uncertainty, students' difficulties from existing studies are summarized into three levels of knowledge integration as follows:

1. Novice level

Novice students' knowledge structures are typically fragmented, with links only connecting the surface features of contextual variables (bottom layer components) and the task goals without a meaningful understanding of the underlying mechanisms. These students can only correctly answer some typical questions based on memorization of learned examples. In addition, due to a lack of a basic understanding of the origin of random uncertainty, novice students often believe that it is unnecessary to make multiple trials unless there is something wrong with a measurement [48]. Although some would make repeated measurements in an experiment, they tend to use a recurring value [49] for the result or simply do it as a routine procedure [47]. The formality of repeated measurement is easily memorized and can be directly applied to similar problems. When working with atypical questions, such as identifying the sources of uncertainty, novices often ascribe all deviations in measurement results to human errors. Additionally, novice knowledge structures may not contain the mechanistic origins of different types of uncertainties, which are rarely discussed explicitly in traditional instruction.

2. Intermediate level

Students at this level develop more connected knowledge structures, allowing them to relate the contextual variables and instruments to the layer of intermediate operations, which connect to the task goals. However, these students' knowledge structures are still fragmented without the integrated understanding that links to the central idea. Therefore, students at this level still solve problems by relying on memorized procedures and examples without clearly considering the central idea. They can perform well on typical questions with familiar contexts but often fail on atypical ones involving novel contexts. For example, students at this level believe it is necessary to make repeated measurements and know to use the arithmetic mean to address uncertainty [33,49]. However, these students have yet to develop a complete understanding of the central idea, which hinders their capacity to distinguish between uncertainties of different origins such as the random uncertainty caused by human operations and the systematic uncertainty due to instrument biases. Without a complete understanding of the central idea, students often have difficulties in identifying the sources and causes of different types of uncertainties, leading to confusion about the need and calculation methods in mathematical data processing. When solving typical textbooklike problems, they sometimes can distinguish between different uncertainties and figure out their sources based on memorized examples. Nevertheless, these students often fail to solve problems with atypical contexts due to the lack of a complete understanding of the central idea.

3. Expertlike level

Students with expertlike understanding can relate the contextual features to the central idea, along with several intermediate processes and related principles, to form a well-connected knowledge structure. As for measurement uncertainty, these students can clearly and explicitly reason with the interactive relations between sources of uncertainty and the influencing factors. The integrated knowledge structure allows the students to recognize the sources and causes of uncertainty in both typical and atypical contexts. Furthermore, they can explicitly distinguish random and systematic uncertainties and understand the data processing strategies to address the different uncertainties. Therefore, these students perform well in solving both typical and atypical problems by applying the central idea, which is the hallmark of a well-connected knowledge structure that diverges from the novices' fragmented structures.

In summary, the conceptual framework can represent students' reasoning pathways and common student difficulties in understanding aspects of measurement uncertainty, which can be further used to categorize students into different levels of knowledge integration. To obtain a quantitative assessment of students' levels of knowledge integration, the conceptual framework is applied to guide the development of a multiple-choice instrument that probes the features of students' knowledge structures, which are discussed next.

TABLE I. Designs of measurement settings targeting single vs multiple observers, devices, and measurements.

Question designs	Questions
Single observer	Q2–4, Q7, Q9–10
Multiple observers	Q5–6, Q8, Q11–15
Single device	Q2–3, Q5–6, Q12–13
Multiple devices	Q4, Q7–8, Q9–11, Q14–15
Single measurement	Q4–7, Q14–15
Repeated measurements	Q2–3, Q8–13

D. Development of the measurement uncertainty test

In this research, a concept test on measurement uncertainty was developed based on the conceptual framework to probe students' levels of knowledge integration with an emphasis on targeting the conceptual understanding of the mechanisms underlying measurement uncertainty. The test contains 15 multiple-choice questions designed with different contexts to target different conceptual elements and the central idea of uncertainty (see the Supplemental Material for the test [51]). To probe different aspects of student reasoning, the questions were designed with three contextual and content configurations including typical and atypical contexts, types of reasoning, and measurement settings to be discussed below.

1. Designs using typical and atypical contexts

For the topic of measurement uncertainty, contextual features often involve different configurations of observers, measurement devices, and measurements performed. The most basic contextual design can involve different numbers of observers, devices, and measurements, which are listed in Table I.

Building off the basic contextual features, combinations of multiple context elements can provide refined settings with typical and atypical questions to probe students' deep conceptual understandings. As shown from previous studies, the question designs using typical and atypical contexts were found to be effective in assessing students' understanding of the central idea and their levels of knowledge integration [17–21]. The definition of typical or atypical questions is based on whether questions of similar contexts have been used in instruction or not. It is also noted that the questions categorized as typical or atypical are based on the instruction received by students involved in this research, which may vary in other education systems. In this study, the participants include a group of college undergraduates who have not taken the college-level laboratory course and

TABLE II. Designs of contextual combinations. Note that *Q1 is a filler question and is not included in data analysis; **Q12 and Q13 are both considered typical since participants in this study are familiar with the contexts.

Design	Configurations on the observers, devices, and measurements	Question contexts	
		Typical	Atypical
Design 1 (D1)	Single observer, single device, single measurement	Q1*	
Design 2 (D2)	Single observer, single device, repeated measurement	Q3	Q2
Design 3 (D3)	Single observer, multiple devices, single measurement	Q4	Q7
Design 4 (D4)	Multiple observers, single device, single measurement	Q6	Q5
Design 5 (D5)	Single observer, multiple devices, repeated measurement	Q10	Q9
Design 6 (D6)	Multiple observers, single device, repeated measurement	Q13, Q12**	
Design 7 (D7)	Multiple observers, multiple devices, single measurement	Q15	Q14
Design 8 (D8)	Multiple observers, multiple devices, repeated measurement	Q8	Q11

a group of high school students who have completed the learning of measurement uncertainty in their high school physics classes. Therefore, these students' understanding of uncertainty is largely based on their learning in high school. In the high school physics instruction on measurement uncertainty, students were explicitly asked to perform tasks that require them to make repeated measurements using a single instrument and computing the mean value. In other words, taking multiple measurements to compute mean value is a familiar, almost automatic procedure for these students. Therefore, the design of typical questions often involves the common task that asks students to "compute the mean value (as the final result)," which is the typical routine procedure taught in the Chinese high school physics curricula to address uncertainty. This task often activates students' memories of similar problems, which can lead them to match memorized solutions without meaningful reasoning.

In contrast, atypical questions are designed with unfamiliar contexts that students rarely encounter in traditional instruction, which often makes the memorization-based strategies nonproductive. To successfully solve atypical questions, students would need to develop a basic understanding of the mechanistic origins of different types of measurement uncertainty. Therefore, the design of atypical questions can directly probe students' understanding of the central idea, which is a signature of achieving an expertlike level of knowledge integration.

The measurement uncertainty test includes eight designs of contextual configurations, summarized in Table II. A pair of typical and atypical questions were designed for each contextual configuration except for D1 and D6. Here, D1 gives a very simple contextual configuration, in which an atypical question cannot be designed. The corresponding Q1 is a simple question, familiar to most students, and is used as a filler question, so it is omitted in data analysis. Meanwhile, students involved in this study were also very familiar with the contextual configuration of D6. It was commonly used in instruction, in which students as a group were explicitly asked to perform tasks that require making repeated measurements using a single instrument. Therefore, Q12 and Q13 are both considered typical questions in this study. It is worth noting that the definition of a typical or atypical question is dependent on instruction and can be defined differently in studies involving different instruction. Altogether, the test includes nine typical questions (Q1, Q3, Q4, Q6, Q8, Q10, Q12, Q13, and Q15) and six atypical ones (Q2, Q5, Q7, Q9, Q11, Q14).

2. Designs targeting different types of reasoning

The reasoning for measurement uncertainty can be categorized into three types of questions. A question can ask students to identify a specific form of uncertainty based on a given task, such as random or systematic, which forms a "what" type of question. A question can also ask students to perform a task, such as selecting a specific reading or device, which forms a "how" type of question. Finally, a question can ask students to find an explanation or a reason for an observed uncertainty, which forms a "why" type of question. These question designs target different thinking pathways that can be useful in determining finer details of students' conceptual understanding.

For the measurement uncertainty test, the questions of different reasoning types are summarized in Table III. Specifically, the what questions ask students to identify what kind of uncertainty can be reduced with the operations mentioned in the question. The how questions ask students to determine the operations that can reduce the uncertainty

TABLE III. Designs of question types.

Question types	Question contexts	Questions
What	Typical	Q1, Q3, Q8, Q10, Q13, Q15
How	Typical and atypical	Q4, Q6, Q9, Q12, Q14
Why	Atypical	Q2, Q5, Q7, Q11

of the reported outcome based on measurements given in the question. The why questions probe if students can recognize the mechanistic origins of different types of uncertainties involved in the given measurements. Notably, the what questions are considered typical since the operations involved focus on making repeated measurements and calculating the mean value. Students had plenty of exposure to this kind of operation in instruction and could memorize the types of uncertainties associated with the operation. Therefore, the what questions can usually be solved using memorization-based strategies. The how questions have mixed categories with some being considered typical, which involve simple calculations of mean values similar to the what questions. On the other hand, some how questions are considered atypical when they involve contexts that require the use of the central idea. The why questions are all considered atypical, since the instructions rarely discussed the mechanistic origins of different uncertainties, and successfully solving this type of question requires a basic understanding of the central idea.

In addition to achieving the research goals of this study, the exploration of question designs using different contextual features and reasoning types can also provide useful information for general assessment and instruction in teaching and learning. Knowing the appropriate types of reasoning and contexts to use in assessment and learning tasks in teaching can greatly aid instructors in effectively teaching this important topic with an aim to help students develop an integrated knowledge structure and deep conceptual understanding.

E. Data collection

Assessment data were collected from a total of 406 students in China, among which there were 247 second-year college students from a large-scale comprehensive university and 159 senior high school students from a high-ranking high school. All students had previously learned the relevant content of measurement uncertainty in their high school physics courses. Students were given 40 min to complete the test.

Interviews were also conducted with 18 volunteers from the same pool of undergraduate students after they completed the concept test. Each interview session lasted approximately 30 min. The purpose of the interviews was to identify the reasoning pathways that students used to answer the questions and to figure out which links in the conceptual framework were being used. During the interviews, students were asked to review the test and explain their answers out loud. Additional follow-up questions were also asked to specifically probe students' understanding of the origins of systematic and random uncertainties.

In data analysis, all scores are scaled to 0-1 for easy comparisons. Students' mean scores on question sets designed with different contextual features and reasoning types were compared using t tests, analysis of variance (ANOVA), and Cohen's d effect size to determine the significant influences on students' performances from varied contextual and reasoning designs. The results were used to determine students' knowledge integration levels and to identify the differences and similarities in students' conceptual understandings.

F. Evaluation of validity and reliability of the test

The validity of the measurement uncertainty test is evaluated in two areas including content validity and measurement validity. The content of the test was designed by a team of experts in physics and physics education including three faculty and three graduate students. The design went through a rigorous cycle of development and revision in multiple iterations of piloting and feedback by additional faculty and graduate students in the research institutions of the authors. The content of the final version of the test has been agreed to be scientifically accurate and valid for measurement by the team of designers and evaluators.

For the measurement validity, 80 college students were interviewed during the piloting-revision phase of the development. The interviews were used primarily to check if the students' responses were aligned with the intended constructs of measurement. An evaluation of the consistency indicates a 94.3% agreement between students' explained understanding of the questions and the intended measurement designs. Additional measurement validity can be further evaluated based on whether the quantitative assessment outcomes are consistent with the expected results of the design. As can be seen from the results presented in the following sections, the quantitative assessment outcomes agree well with the expected outcomes of the design. Therefore, based on the interviews and assessment results, the measurement validity of the test can be sufficiently established.

The reliability of an assessment instrument is often evaluated using the Cronbach's α, which is calculated based on the consistency between test items. This is a valid method when the test is unidimensional but will not produce the intended measure when the test has strong multidimensionality. As discussed earlier, the measurement uncertainty test is designed with multiple dimensions of constructs including contexts, reasoning types, and measurement settings. The multidimensionality is clearly demonstrated by the scree plot included in Fig. 5 in the Appendix, which shows a number of eigenvalues in a similar range. Therefore, Cronbach's α, which produces a small value of 0.18 because of the strong multidimensionality of the text, does not work well in this case.

The original definition of reliability is test-retest consistency. Therefore, we can use the bootstrapping method to simulate the test-retest scenario by resampling the total dataset into multiple subgroups and comparing their means. In this simulation, the total dataset is randomly split into

TABLE IV. Typical and atypical scores and statistical significance of differences.

Contexts	Mean	SD	t	p	Cohen's d
Typical	0.68	0.18	20.07	< 0.001	1.36
Atypical	0.44	0.17			
Total	0.58	0.13			

two subgroups at half of the original size ($N = 203$). Then a t test is conducted to compare the mean scores of the two subgroups. This resampling and comparison are conducted 100 times, which produces an average p value $= 0.51$ and an average Cohen's d effect-size $= 0.08$. The simulation results suggest that the test scores are not statistically significant between test-retest runs with an average uncertainty equivalent to 8% of the standard deviation, which suggests that a satisfactory level of reliability is established.

III. RESULTS

A. Students' performances on different question designs

1. Performances on typical and atypical questions

As shown from previous research, the use of typical and atypical questions in assessment is effective in probing students' understanding of the central idea, which helps to determine their levels of knowledge integration [17–21]. Students' average scores on the typical and atypical questions are given in Table IV, which shows a significant difference between mean scores on the two types of questions [$t(406) = 20.07$, $p < 0.001$, $d = 1.36$].

To investigate fine-grained performance details, students' mean scores on typical and atypical questions in each contextual configuration (see Table II) are also plotted in Fig. 2, which shows that the different designs provide a variety of difficulties and discriminations for assessment of students at a wide range of performance levels. Overall, the designs of typical and atypical questions used in the measurement uncertainty test provide appropriate discrimination and contextual variation to probe students' knowledge integration, which will be discussed in later sections.

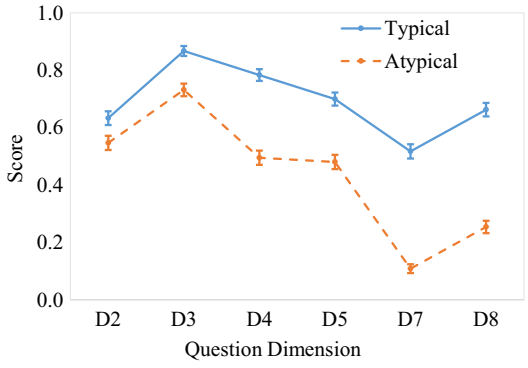

FIG. 2. Students' performance across different designs of contextual configurations. The error bars represent standard errors.

TABLE V. Students' scores on what, how, and why questions.

Question type	Question context	Mean ($N = 406$)	SD
What	Typical	0.66	0.21
How	Typical and atypical	0.57	0.19
Why	Atypical	0.51	0.23

2. Performances on questions targeting different reasoning types

Students' scores on what, how, and why questions are listed in Table V. The assessment outcomes show that students have the highest scores on what questions and the lowest scores on why questions. A one-way ANOVA shows significant differences between the three question types [$F(2, 1218) = 54.85$, $p < 0.001$], which are more clearly demonstrated with pairwise t tests between different question types [$t_{(\text{what-how})}(406) = 6.83$, $p < 0.001$, $d = 0.47$; $t_{(\text{how-why})}(406) = 4.10$, $p < 0.001$, $d = 0.28$; $t_{(\text{what-why})}(406) = 10.38$, $p < 0.001$, $d = 0.70$]. Since many students would rely on memorization-based strategies in problem solving, it is often expected that the what questions are relatively easy for students, whereas the memorization-based approaches can be productive. Meanwhile, the how and why questions are more difficult since they require an increasing level of understanding of the central idea of the concept and cannot be solved with factual memorization.

3. Performances on questions with varied numbers of observers, devices, and measurements

In typical lab activities and assessments, the numbers of observers, devices, and measurements are often varied to create different experimental tasks and problems. Therefore, results on the possible influences of these factors on students' performances can provide valuable information for lab instruction and assessment (see Table VI). The results suggest that students' performances decreased when multiple observers [$t(406) = 12.33$, $p < 0.001$, $d = 0.82$] and measurement devices [$t(406) = 5.51$, $p < 0.001$, $d = 0.39$] were involved. However, changes in the number of measurements did not lead to any significant performance differences [$t(406) = 1.06$, $p = 0.29$)]. These results can be expected since during lab instruction students were frequently asked to conduct repeated measurements in doing an experiment. Although most students do not exactly understand the reason for repeated measurements [33,37,40,41,44,47,48], they are familiar with the context and can solve typical questions correctly using memorization-based strategies. On the other hand, students were

TABLE VI. Students' mean scores on questions involving single or multiple observers, devices, and measurements.

Question designs	Mean	SD	t	p	Cohen's d
Single-observer	0.66	0.18	12.33	<0.001	0.82
Multiple-observers	0.51	0.18			
Single-device	0.62	0.22	5.51	<0.001	0.39
Multiple-devices	0.54	0.17			
Single measurement	0.58	0.19	1.06	0.29	
Repeated measurements	0.57	0.18			

much less familiar with contexts that contain multiple observers or measurement devices. Therefore, increasing the number of observers and/or devices often makes the questions more difficult for students.

B. Quantitative study on students' knowledge structures of measurement uncertainty

To examine how students at different total performance levels may respond to the typical and atypical questions, score distributions of the two types of questions are plotted in Fig. 3. A histogram of the frequency of students' total scores is displayed in the background to show the distribution of students at different performance levels.

As shown in Fig. 3, scores on typical and atypical questions are similarly low for all students with low total scores (score<0.5 marked as 0.4), indicating a novice level of understanding that leads to poor performance on both typical and atypical questions. As the total score increases ($0.5 \leq$ score < 0.9), a performance gap between typical and atypical questions is more pronounced, suggesting that students in this range have started to perform well on typical questions using memorization-based strategies but have not yet developed a good understanding of the central idea. As the total score further improves ($0.9 \leq$ score < 1.0), the performance on typical questions reaches near mastery, and the performance on atypical questions starts to show a noticeable improvement. At this level, students would have developed partially integrated knowledge structures with some understanding of the central idea that allows them to successfully solve most typical questions but still fail on many atypical ones. Finally, students with the highest scores (1.0) display a small difference between their scores on typical and atypical questions, indicating that they have developed a solid understanding of the central idea with a well-integrated knowledge structure. Due to the constraints of the population studied, the number of high-performing students is very small, which also suggests that a good understanding of the central idea is often difficult to achieve in traditional instruction.

To investigate patterns of student reasoning when dealing with measurement uncertainty, the score distributions of questions in different reasoning types including the what, how, and why questions (see Table III) are plotted in Fig. 4. As shown in Fig. 4 and Table V and discussed earlier, the what questions are relatively easy for students since they can be solved with memorization-based strategies. On the other hand, the why questions are much harder, as most students have not yet developed a deeper understanding of the central idea through traditional instruction. The how questions reveal an interesting pattern, which starts to be

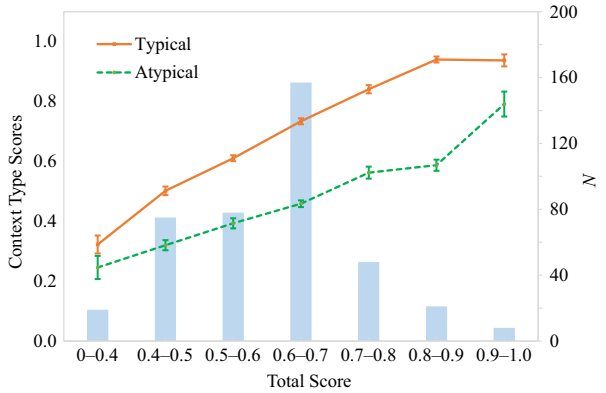

FIG. 3. Plot of typical and atypical questions across total scores (with error bars denoting standard error) for all students in this study. The frequency of total score distribution is shown as a bar chart in the background. An absolute count of students falling into each range is shown on the right axis.

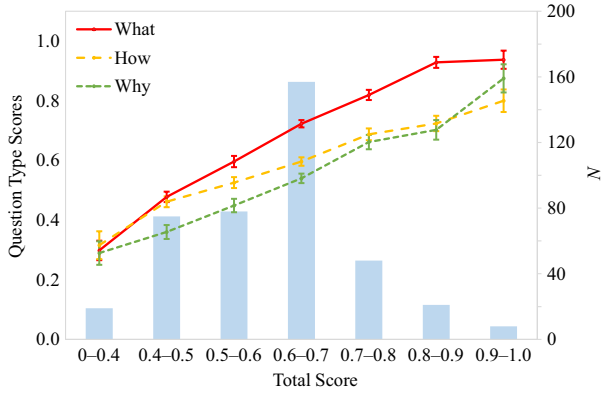

FIG. 4. Plot of what, how, and why questions across total scores (with error bars denoting standard error) for all students in this study. The frequency of total score distribution is shown as a bar chart in the background. An absolute count of students falling into each range is shown on the right axis.

TABLE VII. Summary of total score, question contexts, and question type scores per knowledge integration level. Standard errors are given in parentheses.

Knowledge integration level	Total score	N	Typical	Atypical	What	How	Why
Novice	0.0–0.5	48	0.40 (0.02)	0.29 (0.02)	0.37 (0.02)	0.37 (0.02)	0.34 (0.03)
Intermediate	0.5–0.9	350	0.71 (0.01)	0.45 (0.01)	0.69 (0.01)	0.59 (0.01)	0.52 (0.01)
Expertlike	0.9–1.0	8	0.94 (0.02)	0.79 (0.04)	0.94 (0.03)	0.80 (0.04)	0.88 (0.05)

similar to the what questions for students with low total scores. However, for students with higher total scores, their performance on the how question did not improve with the what questions but, instead became similar to their performance on why questions. A possible reason for this pattern is that traditional instruction usually emphasizes repeated measurements as a necessary procedure but rarely explains why this is needed or how it is connected to measurement uncertainty. Therefore, students often memorize the procedure and can perform well on some typical questions that require repeated measurements, but the strategy will not work on problems that cannot be solved by making repeated measurements. These problems behave more like the why questions which require a deeper understanding of the central idea which was only developed among the very top students.

The general trend of students' score distributions is consistent with their performances on typical and atypical questions shown in Fig. 3. Students with low total scores (score<0.5) performed similarly and poorly among the three types of questions due to their novice level of understanding. As the total score increases (score = 0.5–0.9), the performance gap among the three question types becomes more pronounced, suggesting that students in this range have started to perform well on the what questions using memorization without yet establishing a basic understanding of the central idea which is needed to solve the how and why questions. As the total score further improves (score>0.9), the what question performance is near mastery, and students' performances on the how and why questions start to show significant improvement. Finally, students with a high score near 1.0 show a minor difference among different types of questions which is an indicator that these students have achieved a good understanding of the central idea with a well-integrated knowledge structure.

Based on the gap between typical and atypical questions shown in Fig. 3 and between the different reasoning type questions shown in Fig. 4, a total of three knowledge integration levels can be categorized as shown in Table VII. The score division given in Table VII is identified as the categorization scheme to match between total score and knowledge integration levels. However, since the assessment outcomes and interviews are population dependent, this score division scheme reflects only a reasonable approximation and should not be generally extended to other contexts and populations. Nevertheless, this result demonstrates the possibility of identifying a quantitative categorization scheme to model knowledge integration as well as its utility in teaching and learning. To further validate the knowledge integration levels, interviews were conducted and used as confirmative evidence to support the categorization scheme defined in Table VII. The interview outcomes are discussed next.

C. Qualitative study on students' knowledge structure of measurement uncertainty

In traditional physics instruction, the origins or mechanisms of measurement uncertainty are usually not discussed in detail. This limitation often leads to student difficulties in understanding the causes and interactions within and between systematic and random uncertainties. Typically, students often use a memorization strategy to simply match systematic uncertainty to instruments and random uncertainty to human error.

Since students were unfamiliar with the mechanism of uncertainty, they tended to solve the problems through memorization instead of reasoning. When interviewed, some undergraduate students either did not directly link, or only weakly linked, the cause and solutions, responding with comments such as, "well, there are two kinds of uncertainties, the systematic and random uncertainties, right? …Oops, I forgot how to distinguish them, but I think that making repeated measurements can reduce all types of uncertainty because we do it in class nearly all the time." A few students with higher total scores recognized the connections among the mechanisms and solutions to different types of uncertainties in their responses: "Generally, the observer will cause random uncertainty when reading and recording, we should try to measure as many times as possible and then the average value gets closer to the true value. Meanwhile, systematic uncertainty is usually related to instruments, and we can decrease it only by modifying experimental instruments or perfecting measurement principles." Overall, it is clear that students generally did not correctly or just simply link mechanisms and solutions within their understanding of the different types of uncertainty.

To gain insight into the actual reasoning pathways of students at different performance levels, which are shown in Table VII, interview results are analyzed and discussed next along with the assessment outcomes.

1. Novice level (total score < 0.5*)*

Students performed poorly on all types of problems. When solving problems, these students relied heavily on memorized rules or related real-world intuition. Additionally, these students exhibited little understanding of the central idea and instead directly related elements of the surface features to their responses. Students who exhibited this behavior had thoughts similar to the interview excerpts shown below:

Student A: (Q4) The student chose answer D and explained "I subconsciously regard the choice related to mean value as the best one, because I think the temperature would be high and low, and the average may 'balance' this. Uhh….actually I'm not sure what the 'balance' is and what can be balanced."

Student B: (Q7) The student chose answer D and explained "Although all thermometers are placed in a constant temperature cage, I think it cannot be sure that the temperature is the same everywhere in this temperature cage."

Student C: (Q4) The student chose answer C and explained "I think the thermometers on the shelf are manufactured along with uncertainty and the uncertainty is different from each other. If there are more thermometers giving the same temperature, it indicates that there is no problem in their manufacturing process."

Student D: (Q4) The student chose answer C and explained "Thermometers on the same shelf should not differ too much, so selecting thermometers showing the same temperature would be possible to be close to the exact value… The true value cannot be measured by any method."

As shown from the interviews, when thinking about the solutions to decrease the uncertainty, students at this level often relied on the idea of repetition of outcomes, considering the recurring value as a better value, displaying memorization of the mean value method without reasoning. These students did not understand systematic uncertainty, so they usually solved problems by guessing. Both students A and B mainly attributed the uncertainty to the temperature fluctuation rather than the properties of the measurement devices. Although student A selected the correct answer, the student still misunderstood the sources of uncertainty and was puzzled about the significance of the mean value method. Student C appeared to recognize that the instruments had manufacturing-based systematic uncertainty, but the student did not seem to understand that the manufacturing process can involve random factors that need to be addressed using the mean value method applied to outcomes measured with multiple instruments. There were also novice students who seemed to completely lack the understanding of systematic uncertainty. For example, Student D believed that the instruments should produce identical measures and that choosing ones showing similar outcomes would make the best measurement.

Overall, the novice students performed weakly on most questions. Their problem-solving approaches generally fell into the category of memorization of problems' solutions and rules or guesswork. These poor performances could link to students' fragmented knowledge structures and minimal understanding of the central idea.

2. Intermediate level (total score between 0.5 *and* 0.9*)*

The students at this level exhibited a range of behaviors, but they all had significantly higher mean scores on typical ($S_{\text{typical}} = 0.71$) and atypical ($S_{\text{atypical}} = 0.45$) questions when compared to novices (0.40 and 0.29, respectively) [$t_{\text{typical}}(398) = 14.75$, $p < 0.001$, $d = 2.27$; $t_{\text{atypical}}(398) = 6.74$, $p < 0.001$, $d = 1.04$]. Meanwhile, these students also performed better on what ($S_{\text{what}} = 0.69$), how ($S_{\text{how}} = 0.59$), and why ($S_{\text{why}} = 0.52$) questions than the novices (0.37, 0.37, and 0.34, respectively) [$t_{\text{what}}(398) = 11.60$, $p < 0.001$, $d = 1.79$; $t_{\text{how}}(398) = 8.23$, $p < 0.001$, $d = 1.27$; $t_{\text{why}}(398) = 5.57$, $p < 0.001$, $d = 0.86$].

These students also demonstrated a mixture of using memorization-based strategies and having limited reasoning using the central idea. They often demonstrated inconsistent reasoning depending on the contexts of the questions. For example, student E was able to respond that computing the average of different thermometers' measurement outcomes could reduce their systematic uncertainty in question Q4, but thought that using several rulers to take repeated measurements and computing the mean value could not decrease the systematic uncertainty of rulers:

Student E: (Q4) The student chose answer D and explained "Selecting the thermometer that shows the temperature near the mean value may be better. The systematic uncertainty of the thermometer would be reduced by taking an average value. Well, we always compute the mean value to reduce the uncertainty in the lab course."
(Q10) The student chose answer B and explained "Oh, the systematic uncertainty of each ruler is fixed, right? The division value is the uncertainty of the ruler. Although you used five rulers to measure the height six times, only the random uncertainty of reading can be reduced. The uncertainty of the ruler still cannot be decreased."

(Q9) The student chose answer D and explained "I hesitate because the division value of the ruler is fixed, so I think the systematic uncertainty of each ruler is unchangeable? … Then I thought about whether it still needs to make repeated measurements. It is difficult for me to choose between A and D… The difference between each ruler should be a fixed value, and there is no need to measure it again, so it is enough to make a single measurement and compare the readings of each ruler. I am also afraid that a single measurement will lead to the effect of random readings, but the differences between the rulers must be a fixed value… so, there is no need to measure the height again."

From students' descriptions, the question contexts directly affected reasoning, with student E demonstrating at least partial reasoning using the central idea on the atypical question Q9 and typical question Q10 but relying on matching the surface features with memorization on the typical question Q4. Overall, the intermediate students' reasoning on question Q4 was significantly improved when compared to novices, with these students being able to fig out the cause of the different values of the thermometers.

3. Expertlike level (total score $>$ 0.9)

These students demonstrated near mastery in all typical and atypical questions with the most notable improvement over intermediate students on typical ($S_{\text{typical}} = 0.94$) and atypical ($S_{\text{atypical}} = 0.79$) questions, which are significantly better than those at intermediate level (0.71 and 0.45, respectively) [$t_{\text{typical}}(358) = 9.52,\ p < 0.001,\ d = 3.41$; $t_{\text{atypical}}(358) = 6.07,\ p < 0.001,\ d = 2.17$]. Meanwhile, they also performed better on the what ($S_{\text{what}} = 0.94$), how ($S_{\text{how}} = 0.80$), and why ($S_{\text{why}} = 0.88$) questions than the intermediate students (0.69, 0.59, and 0.52, respectively) [$t_{\text{what}}(358) = 3.72,\ p < 0.001,\ d = 1.33$; $t_{\text{how}}(358) = 3.39,\ p = 0.001,\ d = 1.21$; $t_{\text{why}}(358) = 4.60,\ p < 0.001,\ d = 1.64$]. The expertlike understanding establishes a well-integrated knowledge structure such that students were able to consistently answer typical and atypical questions using the central idea, which is evident from the interview excerpts shown below:

Student F: (Q9) The student chose answer A and explained "Making repeated measurements and calculating the mean value can reduce the random uncertainty for each ruler. Besides, by comparing the mean values of different rulers, we can identify whether systematic uncertainty exists.

Student G: (Q4) The student chose answer D and explained "There are many thermometers, so by computing the mean value, we can reduce the systematic uncertainty from the thermometer. It is consistent with the method in our daily experiment."
(Q7) The student chose answer C and explained "Well, five thermometers, although they are designed to have the same range and division value, there should be some differences among them…that will cause the systematic uncertainty, so the readings for these five thermometers are different."
(Q9) The student chose answer A and explained "To identify whether the rulers are manufactured with different systematic uncertainty, so we need many rulers, but the single measurement is not enough…the best answer should be 'using several rulers to make repeated measurements respectively' (reading the question again) Uhh… there have been five rulers, so the answer is A. By making repeated measurements for each ruler to compute the mean value, we can reduce and eliminate the impact of random uncertainty of reading. If the mean values for each ruler are different, the differences should be caused by the ruler."
(Q10) The student chose answer C and explained "Similarly, there are six measurements per ruler, and computing the average will reduce the random uncertainty of reading. Meanwhile, there are five rulers, so averaging the mean value of each ruler again can then decrease the ruler's systematic uncertainty."

From the interviews, students at this level were able to recognize the mechanisms of different uncertainties and the methods to address them. Both of these students focused their reasoning on the central idea. Contextual factors, such as the condition of the environment and the number of observers and instruments, did not affect their application of the central idea in their reasoning about the mechanism of measurement uncertainty.

According to the interview results, 4 out of the 18 students were identified at the novice level. Their answers were related to surface feature elements and exhibited a high degree of dependence on guesswork. Meanwhile, the majority of the interviewed students (12 out of 18) were identified at the intermediate level. Most of these students directly matched the questions with their memorized procedures before answering the questions, with some students demonstrating some reasoning using the central

idea. These students often exhibited inconsistent reasoning depending on the contexts of the questions, which demonstrates still fragmented knowledge structures. Finally, two students were identified at the expertlike level. They applied the central idea consistently in different contexts to explain the mechanisms of uncertainties and the methods to address the uncertainties. Such problem-solving behaviors demonstrate that these students had developed an integrated knowledge structure, with global and strong links between the central idea and other components of the conceptual framework.

Overall, the assessments and interviews revealed common and persistent difficulties in students' understanding of measurement uncertainty. According to the results from the concept test and interviews, students at different levels of knowledge integration demonstrate unique types of reasoning pathways that can be mapped in the conceptual framework, making it a useful tool for visualizing different reasoning states and knowledge structures.

IV. DISCUSSION AND CONCLUSIONS

In this study, a conceptual framework of measurement uncertainty was developed to guide the assessment of students' knowledge integration in learning. Based on assessment data and interview results, students were categorized into three levels of knowledge integration including novice, intermediate, and expertlike. The reasoning pathways of students at different levels revealed a progression of reasoning from a rudimentary surface level to a deep understanding and can be mapped in the conceptual framework.

Students at the novice level performed poorly on most questions and demonstrated little understanding of the mechanisms of different types of measurement uncertainties. In problem-solving, these students often used memorized procedures, which were directly linked to specific contextual features. As a result, they focused more on the impact of the environment (such as those in their responses to Q4 and Q7). Most of them regarded making repeated measurements to compute the average as an experimental requirement without understanding the underlying mechanisms. Some also held the belief that the recurring value is the correct result, which demonstrates their lack of understanding of systematic uncertainty which has also been reported in previous studies [47–49]. Students at this level were able to solve some simple typical questions using memorized operations or rules but usually failed on more complex typical questions and all atypical questions.

Students in the intermediate level demonstrated better performance than novices on typical questions but had similarly poor performance on atypical questions. They were able to move beyond simple memorization of solutions and demonstrated some basic understanding of the central idea, which was applied in their reasoning on some typical questions. However, these students still failed to apply the central idea to atypical questions and reverted to relying on memorized solutions and procedures. Specifically, these students often applied simple pattern-matching rules, such as considering that instruments have only systematic uncertainty, which cannot be reduced in any way while considering that human observations would lead to random uncertainties that can be reduced by taking an average of repeated measurements. As an important improvement compared to novices, these students did not recognize the causality of measurement uncertainty but exhibited partial reasoning using the central idea of typical questions. However, these students did not exhibit consistent use of the central idea in their reasoning, and their reasoning was often significantly influenced by the contextual features of the questions. The results indicate that the students had fragmented knowledge structures that were still largely memorization based and linked locally to specific contextual features. As a result, these students can answer the typical what questions and some how questions but usually fail on the why questions.

Students at the expertlike level established a more integrated knowledge structure that wraps around the understanding of the mechanisms of measurement uncertainties (the central idea), which allowed them to solve problems successfully with explicit usage of the central idea in different contexts. These students could correctly identify the sources of different types of uncertainties and knew corresponding methods to address them. Furthermore, the central idea was strongly connected to other components within their knowledge structures and was applied consistently to all questions. Students at this level were able to correctly answer all typical questions and most atypical ones.

Results from this study suggest that the conceptual framework of measurement uncertainty is valid in representing and modeling student knowledge structures and assessing student knowledge integration. The assessment outcomes also reveal that the traditional curriculum is not effective in helping students develop an integrated, deep conceptual understanding of measurement uncertainty. To promote knowledge integration in teaching and learning, it is suggested that instructors should emphasize and clearly establish the central idea of measurement uncertainty and develop connections between the central idea and other knowledge components such that the knowledge structure can be activated and trained as an integrated network. In practice, instructors can demonstrate how to solve problems with familiar and novel contexts using the central idea and the connected network of knowledge. In addition, with an established conceptual framework, instructors can know more about the features of the knowledge structures and thinking pathways of students at different knowledge integration levels, which can help better target the missing connections between the central idea and other knowledge components to promote knowledge integration in teaching.

Because of the limited scale of this study, there are a few limitations, which should be further explored in future research. First, the population studied in this research had only a small number of sophomores, which limited the scope of the analysis on students at higher intermediate and expertlike levels. It would be beneficial to study a population with a large number of advanced students so that the developmental progression of knowledge integration on measurement uncertainty can be more thoroughly examined. Meanwhile, the intermediate-level students demonstrated a wide range of reasoning pathways in their interviews, such as identifying different sources of uncertainties in the contexts where observers used different instruments. Additional assessment questions should be developed to target these fine-grained reasoning pathways. It would also be valuable to further investigate students' problem-solving behaviors on questions involving multiple measurements using different instruments so that more reasoning pathways can be examined.

Furthermore, discussions with colleagues also pointed out that it is important to explicitly address two perspectives on measurement uncertainty in a lab course. One is the philosophical understanding that there are always unaccounted factors contributing to an observed measurement uncertainty. The other is the operational method of treating the different possible factors. It is recommended that instructors should discuss both perspectives clearly and frequently in teaching. For example, the measurement instruments described in the questions of the assessment tool of this study may have additional factors contributing to the observed measurement uncertainty, such as possible nonlinearity in the scale marking of a single ruler or a thermometer, or observable spatial temperature differences across a few centimeters of separation in static air in a lab room. Although these factors are fundamentally possible, they are considered in this study to be small-chance events that are often ignored in calculating measurement uncertainty. Therefore, in a real-world lab setting, analysis of observation-based measurement outcomes should always be based on an operational method of decision making through contrast of the probabilities of different contributing factors. In such analysis, the knowledge for distinguishing between small- and large-chance factors in a real-world setting is in fact an essential component of lab skills, which needs to be trained in a lab course. Students need to develop this knowledge so that they can focus on things that make major contributions to modeling measurement uncertainty. The assessment developed in this study can be a useful tool to identify students lacking this knowledge and help deliver effective instruction such as when used as clicker questions.

ACKNOWLEDGMENTS

The authors would like to acknowledge the help of the Editor and anonymous reviewers. The research is supported in part by the Chinese Ministry of Education's Advanced Basic Science Program 2.0 under the Grant No. 20222152 and Guangdong Province Education Research Planning Project (Higher Education Special) of P.R. China under the Grant No. 2022GXJK175, as well as the National Science Foundation Grants No. DUE-2043817 and No. DUE-2110343.

APPENDIX: DIMENSIONAL ANALYSIS OF THE MEASUREMENT UNCERTAINTY TEST

The scree plot shows the eigenvalues in descending order of the correlation matrix of the test. If the test has a unidimensional structure, the first eigenvalue will be much larger than the second eigenvalue. If the test has a strong multidimensional structure, the first eigenvalue will be comparable to the second and additional eigenvalues.

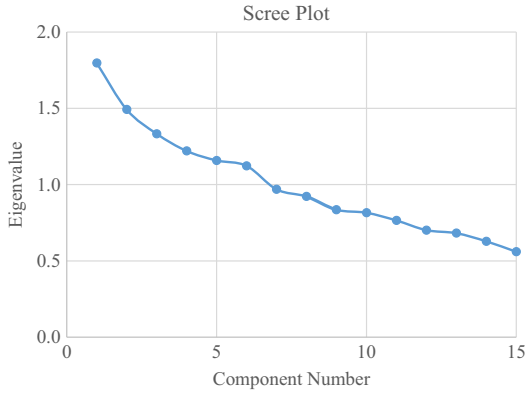

FIG. 5. Scree plot of the eigenvalues of the correlation matrix of the measurement uncertainty test.

[1] L. Bao and K. Koenig, Physics education research for 21st century learning. Discip. Interdiscip. Sci. Educ. Res. **1**, 2 (2019).

[2] E. Kim and S.-J. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, Am. J. Phys. **70**, 759 (2002).

[3] P. G. Hewitt, Millikan Lecture 1982: The missing essential—a conceptual understanding of physics, Am. J. Phys. **51**, 305 (1983).

[4] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, Phys. Teach. **33**, 138 (1995).

[5] P. V. Engelhardt and R. J. Beichner, Students' understanding of direct current resistive electrical circuits, Am. J. Phys. **72**, 98 (2004).

[6] E. Gaigher, J. M. Rogan, and M. W. H. Braun, Exploring the development of conceptual understanding through structured problem solving in physics, Int. J. Sci. Educ. **29**, 1089 (2007).

[7] P. Nieminen, A. Savinainen, and J. Viiri, Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning, Phys. Rev. ST Phys. Educ. Res. **8**, 010123 (2012).

[8] A. Elby, Another reason that physics students learn by rote, Am. J. Phys. **67**, S52 (1999).

[9] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).

[10] A. E. Rivet and J. S. Krajcik, Contextualizing instruction: Leveraging students' prior knowledge and experiences to foster understanding of middle school science, J. Res. Sci. Teach. **45**, 79 (2008).

[11] D. R. Krathwohl, A revision of Bloom's taxonomy: An overview, Theory Pract. **41**, 212 (2002).

[12] NGSS Lead States, *Next Generation Science Standards: For States, by States* (National Academies Press, Washington, DC, 2013).

[13] M. C. Linn, The knowledge integration perspective on learning and instruction, in *The Cambridge Handbook of the Learning Sciences*, edited by K. Sawyer (Cambridge University Press, New York, 2006), pp. 243–264.

[14] M. S. Sabella and E. F. Redish, Knowledge organization and activation in physics problem solving, Am. J. Phys. **75**, 1017 (2007).

[15] O. L. Liu, H. S. Lee, and M. C. Linn, Measuring knowledge integration: Validation of four-year assessments, J. Res. Sci. Teach. **48**, 1079 (2011).

[16] J. L. Snyder, An investigation of the knowledge structures of experts, intermediates and novices in physics, Int. J. Sci. Educ. **22**, 979 (2000).

[17] R. Dai, J. C. Fritchman, Q. Liu, Y. Xiao, H. Yu, and L. Bao, Assessment of student understanding on light interference, Phys. Rev. Phys. Educ. Res. **15**, 020134 (2019).

[18] Y. Nie, Y. Xiao, J. C. Fritchman, Q. Liu, J. Han, J. Xiong, and L. Bao, Teaching towards knowledge integration in learning force and motion, Int. J. Sci. Educ. **41**, 2271 (2019).

[19] W. Xu, Q. Liu, K. Koenig, J. Fritchman, J. Han, S. Pan, and L. Bao, Assessment of knowledge integration in student learning of momentum, Phys. Rev. Phys. Educ. Res. **16**, 010130 (2020).

[20] L. Bao and J. C. Fritchman, Knowledge integration in student learning of Newton's third law: Addressing the action-reaction language and the implied causality, Phys. Rev. Phys. Educ. Res. **17**, 020116 (2021).

[21] L. Xie, Q. Y. Liu, H. Lu, Q. Y. Wang, J. Han, X. M. Feng, and L. Bao, Student knowledge integration in learning mechanical wave propagation, Phys. Rev. Phys. Educ. Res. **17**, 020122 (2021).

[22] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, Science **208**, 1335 (1980).

[23] M. T. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[24] M. T. H. Chi, P. J. Feltovich and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[25] L. Bao, K. Hogg, and D. Zollman, Model analysis of fine structures of student models: An example with Newton's third law, Am. J. Phys. **70**, 766 (2002).

[26] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).

[27] M. F. Fox, A. Werth, J. R. Hoehn, and H. J. Lewandowski, Teaching labs during a pandemic: Lessons from Spring 2020 and an outlook for the future, arXiv:2007.01271.

[28] American Association of Physics Teachers, Goals of the introductory physics laboratory, Am. J. Phys. **66**, 483 (1998).

[29] T. S. Volkwyn, S. Allie, A. Buffler, and F. Lubben, Impact of a conventional introductory laboratory course on the understanding of measurement, Phys. Rev. ST Phys. Educ. Res. **4**, 010108 (2008).

[30] B. Pollard, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab on students' understanding of measurement uncertainty, arXiv:1707.01979.

[31] H. J. Lewandowski, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and B. Pollard, Student reasoning about measurement uncertainty in an introductory lab course, arXiv:1707.01980.

[32] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. De Vries, What do they know? Investigating students' ability to analyse experimental data in secondary physics education, Int. J. Sci. Educ. **43**, 274 (2021).

[33] D. L. Deardorff, Introductory physics students' treatment of measurement uncertainty, PhD thesis, North Carolina State University, 2001.

[34] R. Fairbrother and M. Hackling, Is this the right answer?, Int. J. Sci. Educ. **19**, 887 (1997).

[35] J. Leach, R. Millar, J. Ryder, M.-G. Séré, D. Hammelev, H. Niedderer, and V. Tselfes, Survey 2: Students' images of science as they relate to labwork learning. Working paper 4, Labwork in Science Education. European Commission: Targeted Socio-Economic Research Programme, Project PL 95–2005 (1998).

[36] M.-G. Séré, R. Journaux, and C. Larcher, Learning the statistical analysis of measurement errors, Int. J. Sci. Educ. **15**, 427 (1993).

[37] J. Day and D. Bonn, Development of the concise data processing assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).

[38] N. G. Holmes and C. E. Wieman, Assessing modeling in the lab: Uncertainty and measurement, in *Proceedings of the 2015 Conference on Laboratory Instruction Beyond the First Year, College Park, MD* (2015), pp. 44–47.

[39] H. Eshach and I. Kukliansky, Developing of an instrument for assessing students' data analysis skills in the

undergraduate physics laboratory, Can. J. Phys. **94**, 1205 (2016).

[40] A. Susac, A. Bubic, P. Martinjak, M. Planinic, and M. Palmovic, Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **13**, 020125 (2017).

[41] N. Majiet and S. Allie, Student understanding of measurement and uncertainty: Probing the mean, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Majiet.

[42] M. M. Stein, C. White, G. Passante, and N. G. Holmes, Student interpretations of uncertainty in classical and quantum mechanics experiments, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Stein.

[43] E. M. Stump, C. L. White, G. Passante, and N. G. Holmes, Student reasoning about sources of experimental measurement uncertainty in quantum versus classical mechanics, arXiv:2007.06675.

[44] S. Pillay, A. Buffler, F. Lubben, and S. Allie, Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics laboratory, Eur. J. Phys. **29**, 647 (2008).

[45] M. Parappilly, C. Hassam, and R. J. Woodman, Race to improve student understanding of uncertainty: Using LEGO race cars in the physics lab, Am. J. Phys. **86**, 68 (2018).

[46] B. Pollard, A. Werth, R. Hobbs, and H. J. Lewandowski, Impact of a course transformation on students' reasoning about measurement uncertainty, Phys. Rev. Phys. Educ. Res. **16**, 020160 (2020).

[47] A. Buffler, S. Allie, and F. Lubben, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).

[48] R. L. Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, Nord. Stud. Sci. Educ. **2**, 40 (2006).

[49] F. Lubben and R. Millar, Children's ideas about the reliability of experimental data, Int. J. Sci. Educ. **18**, 955 (1996).

[50] S. Q. Guo and W. Q. Li, *Experiment Tutorial in General Physics* (Higher Education Press, Beijing, 2015), 2nd ed.

[51] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.020145 for test of measurement uncertainty.