Introductory physics lab instructors' perspectives on measurement uncertainty

Benjamin Pollard,[1,2,*] Robert Hobbs,[3] Rachel Henderson,[4] Marcos D. Caballero,[4,5,6] and H. J. Lewandowski[1,2]

[1]*Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA*
[2]*JILA, National Institute of Standards and Technology and the University of Colorado, Boulder, Colorado 80309, USA*
[3]*Department of Physics, Bellevue College, Bellevue, Washington 98007, USA*
[4]*Department of Physics & Astronomy and CREATE for STEM Institute, Michigan State University, East Lansing, Michigan 48824, USA*
[5]*Department of Computational Mathematics, Science, & Engineering, Michigan State University, East Lansing, Michigan 48824, USA*
[6]*Department of Physics and Center for Computing in Science Education, University of Oslo, 0315 Oslo, Norway*

(Received 15 February 2021; accepted 5 April 2021; published 6 May 2021)

Introductory physics lab courses serve as the starting point for students to learn and experience experimental physics at the undergraduate level. They often focus on measurement uncertainty, an essential topic for practicing physicists and a foundation for more advanced lab learning. As such, measurement uncertainty has been a focus when studying and improving introductory physics lab courses. There is a need for a research-based assessment explicitly focused on measurement uncertainty that captures the breadth of learning related to the topic, and that has been developed and documented in an evidence-centered way. In this work, we present the first step in the development of such an assessment, with the goal of establishing the breadth and depth of the domain of measurement uncertainty in introductory physics labs. We conducted and analyzed interviews with introductory physics lab instructors across the US, identifying prevalent concepts and practices related to measurement uncertainty, and their level of emphasis in introductory physics labs. We find that instructors discuss a range of measurement uncertainty topics beyond basic statistical ideas like mean and standard deviation, including those connected to modeling, another lab learning goal. We describe how these findings will be used in the subsequent development of the assessment, called the Survey Of Physics Reasoning On Uncertainty Concepts In Experiments (SPRUCE).

DOI: 10.1103/PhysRevPhysEducRes.17.010133

I. INTRODUCTION

Measurement uncertainty is an important learning outcome for physics lab courses, in particular at the introductory level. It is a central component of guidelines for physics lab courses [1], and it supports other common learning goals for labs, such as experimental modeling [2], critical thinking [3], and epistemology and the nature of science [4–6]. Accordingly, measurement uncertainty is a critical topic for practicing physicists to have mastered regardless of subdiscipline [7].

As it is such a central focus of physics lab learning, measurement uncertainty has been a long-standing focus of physics lab educators [8–12] and education researchers [13–19]. For such work, it is of central importance to be able to measure learning around measurement uncertainty. A common and useful way of measuring learning is using research-based assessment instruments (RBAIs) [20]. RBAIs related to measurement uncertainty have been developed, notably the Physics Measurement Questionnaire (PMQ) [21], the Concise Data Processing Assessment (CDPA) [22], and parts of both the Laboratory Data Analysis Instrument (LDAI) [23] and the Physics Lab Inventory of Critical thinking (PLIC) [24]. However, these RBAIs centered around measurement uncertainty are relatively narrow in scope and are missing the full breadth of important concepts and practices in the topic. There is a need for a new RBAI to measure a broader range of learning related to measurement uncertainty, based on

*benjamin.pollard@colorado.edu, he/him/his

well-defined theoretical constructs and created using a development process that is documented thoroughly to connect constructs to design choices through well-reasoned rationales [25]. It is also critical that this RBAI be developed in consultation with relevant stakeholder populations in order to capture the breadth of learning related to measurement uncertainty.

A process for developing RBAIs in this way is known as evidence centered design (ECD) [26]. ECD specifies a series of steps to ensure that the resulting survey measures constructs that are relevant to its stakeholders. In this context, those stakeholders include physics lab instructors, education researchers, students, and practicing physicists. In short, ECD is a development process that establishes the construct validity of an RBAI. In enacting such a process, it is critical to document and publish the results of this development process to provide evidence of this validity [27]. Recently, in physics education, such a development process has been documented for other RBAI development efforts [28–33]. We describe ECD in more detail below in Sec. II C.

In this work, we describe the first step of ECD development, domain analysis, which is situated in a larger project to create an RBAI for measuring student learning related to measurement uncertainty in introductory physics lab courses. Our RBAI will be called the Survey Of Physics Reasoning On Uncertainty Concepts In Experiments (SPRUCE). The goal of this first step is to establish the scope of SPRUCE, acting as a starting point of the central argument of the assessment. This assessment argument provides the rationale for why the results from SPRUCE should be trusted. It, in turn, serves as the starting point for the creation of the survey itself. For SPRUCE to prove useful for widespread use, it must align with the needs of the stakeholders most central to this RBAI. In our context, the most central stakeholders for introductory physics lab courses are introductory physics lab instructors [28]. These instructors have extensive expertise concerning lab learning from their own teaching experience. Therefore, we focus on introductory physics lab instructors' perspectives related to measurement uncertainty, and aim to match the scope of SPRUCE to these perspectives. In future work concerning later stages of the development of SPRUCE, we will get input from additional stakeholders, in particular physics students, as well as soliciting further feedback from physics lab instructors.

To guide the work presented here, we focus our investigation on three research questions:

- RQ1: According to physics lab instructors, what concepts are relevant for understanding measurement uncertainty?
- RQ2: According to physics lab instructors, what practices do students engage with related to measurement uncertainty in lab courses?
- RQ3: According to physics lab instructors, which of these concepts and practices are emphasized in introductory physics labs, and in what contexts?

While concepts and practices are often intertwined, we separate them in this work to foreground the variety of learning modalities and outcomes involved in lab learning. We use "concepts" to refer to abstract factual knowledge, such as traditional physics content from a lecture or textbook. We also include epistemologies related to the nature of science when we refer to concepts, for example, the idea that every number has an uncertainty. In contrast, we use "practices" to refer to how such knowledge is used in the context of an activity or experiment, including particular skills, procedures, and competencies associated with experimental tools and apparatus. For example, "standard deviation" as a concept has to do with its mathematical definition and how it represents the spread of a distribution. As a practice, "calculating standard deviations" involves the process of producing a quantity representing the standard deviation of a dataset, perhaps using software, a calculator, or simply with pen and paper.

To address these research questions, we contacted and interviewed introductory physics lab instructors across a range of institutions in the US. We asked about the concepts and practices related to measurement uncertainty and how they appear in instructors' classrooms. Our analysis of these interviews identified the concepts and practices common across these instructors and which were most prevalent. It also showed the context in which measurement uncertainty was learned, taught, and emphasized.

II. BACKGROUND

A. Perspectives on student learning of measurement uncertainty in physics

As a central physics lab practice, how students learn measurement uncertainty has been the focus of physics education researchers. Here, we provide a brief overview of the models, perspectives, and findings that have been developed to understand student learning of measurement uncertainty in the physics lab classroom. Given its foundational nature, measurement uncertainty appears in most physics lab curricula; therefore, we do not catalog all teaching approaches and assessment instruments that involve measurement uncertainty. We also restrict our scope to experimental physics and closely related disciplines, and leave more abstract questions about measurement to the fields of epistemology and nature of science.

A notable perspective on the learning of measurement uncertainty comes from the work of Allie, Buffler, Campbell, Volkwyn, and others at the University of Cape Town, ZA. This work occurred in conjunction with the development and early use of the PMQ [21] in the context of a first-year physics course. Extending and adapting previous work with primary school students [34], these researchers

developed a model to understand and categorize students' reasoning around measurement uncertainty. This model centers around two paradigms, point and set, pertaining to the statistical uncertainty of measured quantities. These paradigms have been described extensively elsewhere [18,19,21,35]. In short, the point paradigm holds that a single measurement, if performed perfectly, can yield the true value with no uncertainty, while the set paradigm requires that multiple trials be carried out to estimate the true value with decreasing, but ever-present, uncertainty. Recent work by Majiet and Allie investigates students shifting from point to set paradigm as a result of actual conceptual change versus mere recognition of familiar situations [36].

These paradigms have proved to be useful in introductory physics lab learning beyond Cape Town, including in studies in the US [18,19,37,38]. It should be noted, however, that the point and set paradigms concern only a subset of topics related to measurement uncertainty. They focus on statistical uncertainty and the notion of distributions, and do not encompass systematic uncertainty, instrument precision, and many other skills and practices involved in measurement uncertainty in physics labs [1]. For example, Séré *et al.* studied conceptions about measuring in first-year university students [39]. In addition to identifying student struggles involving confidence intervals and the need to make several measurements, their findings aligned with the focus of the point and set paradigms. They also found that students struggle with the distinction between statistical and systematic errors.

More recently, studies around measurement uncertainty have further broadened in scope, investigating systematics and sources of uncertainty, representations, accuracy vs precision, and uncertainty's role in other physics learning. For example, studies involving the Investigative Science Learning Environment (ISLE), a curriculum and learning environment, focus on supporting students to identify sources of uncertainty and to compare results using uncertainties [40]. Susac *et al.* investigated how students' understanding of measurement uncertainty, specifically data processing and data comparison, is affected by the graphical representation of measurement results [41]. Kok *et al.* suggest that among middle school students, a lack of knowledge about measurement uncertainty results in them mistaking highly precise numerical results for accurate or correct measurements [42]. Leak *et al.* document the importance of measurement uncertainty in the optics workforce, identifying it as a central component of the "number sense" that is critical on the job [43]. As further indication of a renewed focus on measurement uncertainty in physics education, Serbanescu and Harrison identify experimental uncertainty as a threshold concept in physics, and call for more qualitative and mixed methods studies on the topic [44,45].

A particularly common focus when connecting measurement uncertainty to physics learning more broadly is to consider its role in the overall process of modeling in experimental science. For example, Masnick, Klahr, and Knowles investigate the development of scientific reasoning in childhood, exploring how beliefs about data variability and consistency affect the ability to revise one's model of a pendulum [46]. In later stages of learning, Hu and Zwickl found differences in approaches to uncertainty between introductory university students and Ph.D. students [47], showing that introductory students used uncertainty representationally to describe imperfections and variability while Ph.D. students viewed uncertainty analysis in an inferential role to inform the experimental (modeling) process. Holmes and Wieman also studied uncertainty in connection to modeling [48]. They explored physical interpretations of measurement uncertainty, rather than focusing on statistical concepts, and found that students conflate measurement uncertainty, systematic effects, and measurement mistakes. Uncertainty is integral in the mode breaking of Vonk *et al.*, with practices such as estimating and propagating uncertainty and evaluating agreement operationalized in their model-breaking rubric [49]. As a final example, Hull *et al.* reviewed student misconceptions in understanding probability across physics topics including measurement uncertainty, and identified an underlying idea from mathematics education that randomness "is incompatible with predictions and laws [50]." These findings suggest that students would benefit from the integration of uncertainty and modeling offered by lab learning.

B. Connection to modeling

A framework that describes experimental modeling, and includes measurement uncertainty topics implicitly, is the experimental modeling framework (EMF) [2]. The EMF describes the process of modeling in the context of a physics experiment, and is a common learning goal in physics lab courses, especially at the upper-division undergraduate level. Via iteration through subtasks, such as constructing models, making comparisons between data and model predictions, proposing causes, and enacting corresponding revisions, the EMF aims to refine models and apparatus until sufficient agreement between data and prediction is achieved. While the EMF does not foreground measurement uncertainty, it contains it implicitly in the make comparisons and propose causes subtasks. Concepts and practices related to distributions, spread, instrument precision, and making numerical comparisons are required in order to determine if data and prediction are in good enough agreement in the make comparisons subtask. If there is not sufficient agreement, concepts and practices related to systematic uncertainty, sensitivity analysis and propagation of error, and identifying sources of uncertainty are central to the propose causes subtask. Thus, measurement uncertainty serves as a foundation for the EMF,

suggesting a progression from introductory-level topics to the full EMF when applied at the upper-division level.

In summary, measurement uncertainty appears as an integral part of physics lab learning. As such, it is often a component of broader investigations in experimental physics education research, for example, concerning workforce needs, or studying student understanding of modeling or data representation. However, measurement uncertainty is often treated in these studies as a single monolithic topic, lacking the specificity and scope of a focused investigation into the relevant subtopics and underlying constructs that make up measurement uncertainty. Even in studies that focus specifically on measurement uncertainty, they often take a narrow view of the topic, for example, focusing only on statistical concepts. In creating an RBAI on measurement uncertainty for introductory physics labs, there is a need for an investigation into the full range of ideas that are invoked when studying, learning, and employing measurement uncertainty.

C. Evidence centered design

Evidence centered design is a theoretically grounded framework for constructing assessments "in terms of evidentiary arguments" that can be used with a wide variety of learning theories [26,51]. ECD is a widely used assessment framework, particularly in K–12 science education [52,53] and in the development of standardized assessments [54]. More recently, it has been used in chemistry education research to build assessments of science practice [33] and to adapt assessment tasks to probe three-dimensional learning [55]. In fact, Stowe and Cooper argue that not using ECD limits what chemistry education researchers can learn about what students know and how students make use of their knowledge [55,56].

Evidence centered design offers a set of 5 layers that facilitate the construction of assessment items: *Domain Analysis*, *Domain Modeling*, *Conceptual Assessment Framework*, *Assessment Implementation*, and *Assessment Delivery.* We provide a brief overview of each layer, and refer the reader to Mislevy *et al.* for more details [26]. In the domain analysis, researchers construct a thematically organized and prioritized list of knowledge and practices that should be assessed in contexts that are meaningful. For our work, we focus on the knowledge, practices, and contexts relevant to measurement uncertainty in introductory physics. Once the domain has been established, researchers can proceed to domain modeling where narrative assessment arguments are constructed that document how proposed tasks and work products have the potential to elicit different forms of understanding. It is here where the construction of evidentiary arguments begins by making claims about what knowledge and practices are being assessed and how evidence of understanding will be demonstrated by students. For us, those claims would focus on student understanding of measurement uncertainty. After those narrative arguments have been constructed and vetted, researchers proceed with developing the conceptual assessment framework. This is the stage that many in physics education research would consider "assessment development" as it is where assessment tasks are designed and piloted and measurement models (e.g., classical test theory [57]; item response theory [58]) are applied to student work. For our work, this consists of developing appropriate assessment tasks meant to elicit student understanding of measurement uncertainty in relevant contexts and testing scoring methods for those tasks. Through this process, we will develop claims and counter-claims regarding the evidence of what forms of understanding our assessment tasks can provide. The assessment implementation layer follows and is used to collect broader data on student understanding through a piloted set of assessment tasks. By collecting this broader data, researchers can validate their measurement models, provide initial reports to stakeholders, and revise the assessment tasks in light of the issues uncovered in the process. For us, this work involves porting our pilot assessment tasks to a web framework, asking lab instructors to deliver these tasks to their students, generating analyses of the pilot to deliver to instructors, and obtaining feedback. Finally, the assessment delivery layer is reached where a complete, validated assessment is produced. In addition, the associated measurement models and reports to stakeholders have been finalized. For our work, this layer results in the final version of SPRUCE and the associated web framework that delivers SPRUCE, scores student responses, and reports results to instructors.

In contrast to assessment development that has historically been conducted in physics education research, ECD makes clear what specific knowledge and practices are being assessed, why certain contexts are used in the assessment, and why those items provide evidence of student learning of that knowledge and those practices in that context. In addition, all aspects of the development are carefully documented and evidentiary claims are developed throughout, which are tested repeatedly. As such, the structure and transparency of ECD as an assessment framework makes clear what claims about understanding can be made and what claims cannot. For the work presented here, we focus on the initial step of domain analysis that establishes "what knowledge is constructed"; "how that knowledge is used"; and "what evidence stakeholders will accept." To do this, we conducted and coded interviews with 22 physics lab instructors, which we describe below.

III. METHODS

A. Recruitment

Physics lab instructors were contacted via email to participate in this study. Instructors were selected by sampling from several databases: the 2018 Carnegie

TABLE I. The number of institutions from which individuals were contacted (first column) and interviewed (second column). All classifications were based on the 2018 Carnegie database ("Selectivity" from the Ugrad profile variable).

Characteristic	Contacted $N = 116$	Interviewed $N = 22$
Highest degree offered		
Ph.D.	36	6
Master's	26	7
Bachelor's	31	6
Two-year	23	3
Selectivity		
More selective	32	10
Selective	34	6
Inclusive	18	1
Not reported	32	5
Student body classification		
Historically Black College or University	7	2
Hispanic Serving Institution	29	3
Minority Serving Institution	49	7
Tribal College or University	7	0
Women's College	14	1

TABLE II. Self-reported information about the 22 instructors whom we interviewed, and the lab course(s) that they teach. Gender and Race or ethnicity categories emerged from the responses of these instructors. Not all instructors answered every question, and some answered with responses that fell into multiple categories.

Characteristic	Interviewed $N = 22$
Gender	
Any other gender	0
Female	8
Male	13
Race or ethnicity	
Latin American	1
Mixed or Multiple identities	2
South Asian or Indian	4
White or Caucasian	14
Type of courses they teach	
Algebra-based	11
Calculus-based	14
Students per term	
<25	4
25–49	0
50–99	7
≥ 100	8
Relationship to lecture	
Coordinated	12
Integrated	5
Stand-alone	4

Classification of Institutions of Higher Education database [59], a roster of physics departments from the American Institute of Physics [60], and a database compiled for a previous project around modeling [28]. Random sampling of the entire database of about 4350 institutions was augmented by sampling from subsets of institutions along particular characteristics in order to obtain a breadth of perspectives of lab instruction across the US. Those characteristics included the highest level of degree offered at the institution (bachelors, master's, and Ph.D.), minority serving institutions, women's colleges, and two-year colleges. We randomly selected 15 institutions each from the three degree level categories, and augmented that list with 30 distinct minority serving institutions, 14 distinct women's colleges, and 27 two-year colleges. This resulted in a total of 116 institutions from which we solicited interview participants. When necessary, institutional and department websites were consulted to identify instructors and physics department chairs to email. These recipients were asked to assist in identifying the most appropriate instructor in their department to interview about introductory physics lab instruction, and were asked to suggest another instructor to contact if they were not the best person to interview.

In total, 217 individuals were contacted to participate in this study. Numbers of these institutions across various characteristics are shown in the first column of Table I. From those contacts, 22 individuals were interviewed in the summer of 2019, each from a different institution. A breakdown of these 22 institutions across the same characteristics is shown in the second column of Table I. Information about those instructors, and the lab course(s) that they teach, was solicited during the interviews and is tabulated in Table II. In Table II, in addition to instructor and basic course characteristics (i.e., algebra-based vs calculus-based; number of students), we have also included the lab course's "relationship to lecture," which describes how the instructor discussed the role of the lab course in relation to any associated theory-focused course. A *coordinated* lab course is one that is associated with a theory-focused course and roughly follows the content of that course, but meets separately from that course. An *integrated* lab course is a course that is combined with theory content, with class sessions alternating or combining theory- and lab-oriented activities. Lastly, a *stand-alone* lab course is a course that is not associated with a theory-focused course, other than perhaps requiring such a course(s) as a pre- or corequisite.

B. Interview format

Semistructured interviews were conducted via video conference, using an interview protocol designed to solicit relevant information for future assessment development. The protocol was divided into four parts: (i) questions about the context of the introductory lab courses that the instructor taught and their department, (ii) the concepts and

practices that are relevant to measurement uncertainty, (iii) how measurement uncertainty is taught and assessed, and (iv) wrap-up questions to collect demographic information. After identifying the courses that the instructor had experience teaching, and the place of those courses in the overall curriculum of the department, the instructor was asked to focus on the particular introductory lab course or courses with which they were most familiar for the remainder of the interview.

The full protocol is reproduced in the Supplemental Material [61]. Each part consisted of several questions, however, the precise order and inclusion of each question deviated slightly from interview to interview based on the instructor's responses and ensuing conversation. Each interview was audio and video recorded for analysis. Interviews lasted between 30–120 min, with most lasting about an hour. These interviews resulted in a total of about 29 h of video data.

C. Interview analysis

Interviews were transcribed using an automated service [62], and then manually corrected. They were then coded in two passes, as illustrated in Fig. 1, with the development of an emergent code book occurring in between the two passes. Below we describe our interview analysis process.

1. First pass

In the first pass, an *a priori* coding scheme was used to classify easy-to-categorize features within the entire data set that required only literal interpretation. They were *a priori* in the sense that their definitions did not emerge from the dataset in this work, rather, they were based on the knowledge and expertise of the authors before interviews were conducted. These *a priori* codes corresponded to particular questions asked in the protocol. They concerned the characteristics listed in Table II. There also were codes pertaining to particular topics covered in the course(s), and particular ways that measurement uncertainty was taught in the course(s). Those codes are listed below in Sec. IV, Table III. These *a priori* codes are represented in Fig. 1 by the green box in the top right corner. As they required no further analysis in the second pass, they are not contained in any of the blue boxes.

Additionally, we coded excerpts in this first pass with top-level codes, saving them for further analysis using emergent subcodes in the second pass. These top-level codes, identified in this section with italic text, represented different categories related to measurement uncertainty. Any excerpt from the interview transcripts that fit into one of these top-level codes was coded in the first pass. These top-level codes are represented in Fig. 1 by the green boxes that are contained in blue boxes. The green boxes are the top-level codes themselves, and the blue box(es) that contain them pertain to the further analysis that we did in the second pass (described below).

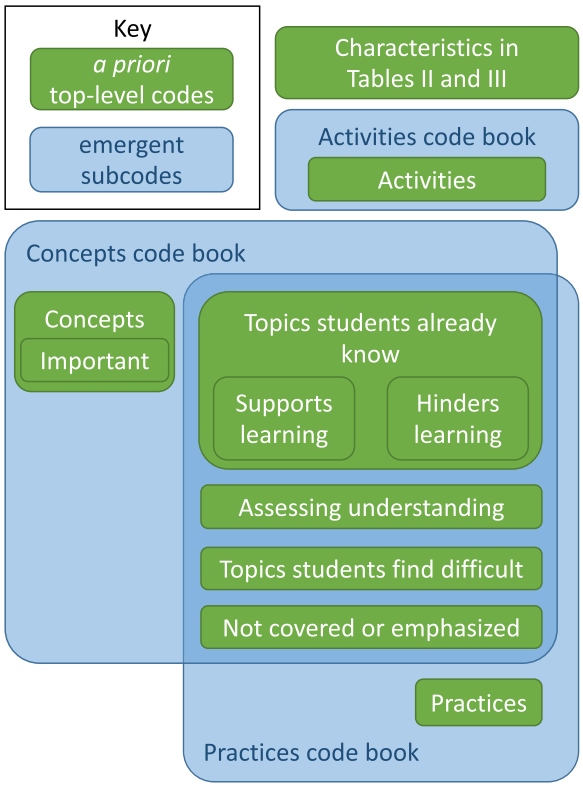

FIG. 1. A diagram illustrating our coding scheme. The green boxes represent the types of *a priori* codes that were applied in the first pass. These codes were used to identify excerpts for later analysis. The blue boxes represent the emergent codes that were applied in the second pass. In that pass, the excerpts pertaining to each green box were coded using the emergent code book(s) of the blue box(es) in which they appear in this diagram.

Three types of excerpts were coded: those pertaining to measurement uncertainty *concepts*, to measurement uncertainty *practices*, and to descriptions of particular lab

TABLE III. The number of interviews to which each of the first-pass *a priori* codes were assigned, pertaining to the particular topics covered in the course(s) and how measurement uncertainty is taught in the course(s).

Code	Count $N = 22$
Topic	
Fitting and significance	22
Systematic uncertainty	19
Propagation of error	18
Significant figures	17
Standard deviation vs standard error	10
Normal vs Poisson distributions	2
How it is taught	
Lab guides	21
Problem sets, homework, or pre-lab activities	15
Textbook	10
Exams	9
Lectures	9

activities. Certain concepts were identified by instructors as particularly *important*; those were additionally assigned this top-level code to mark them as such. Similarly, certain concepts and practices were identified as a way of *assessing* understanding of measurement uncertainty; those were assigned this additional top-level code.

After a broad discussion of concepts and practices associated with measurement uncertainty, instructors were asked more directed questions related to the level of difficulty of these concepts and practices. These excerpts were coded with the following top-level codes, as applicable: *topics students find difficult* and *topics students already know*. Topics students already know were additionally coded when instructors mentioned that these topics either *supported* or *hindered* learning in their course. Lastly, we coded excerpts in which instructors mentioned *topics that they chose not to cover or emphasize* in their course(s).

2. Development of emergent codes

We created emergent code books of subcodes using the excerpts coded with top-level codes in the first pass. While the first pass considered the entire transcripts of all of the interviews, the second pass focused only on the particular excerpts that had been coded with the top-level codes in the

TABLE IV. The set of emergent codes pertaining to measurement uncertainty concepts. The first column is the name of the code, the second is a description of that code, and the third is an example excerpt from the interviews.

Code	Description	Sample quote
Curve fitting	Ideas related to fitting functions to data.	I teach them about the root mean square error, RMSE, and also the correlation coefficient for a linear fit.
Distributions, in general	The need for multiple measurements, the concept of spread, the idea of random error.	If I make the same measurement seven times, I'm going to get the same number seven times? [No, I won't.]
Every number has an uncertainty	The idea that every measured quantity in lab has an associated uncertainty.	That uncertainty is an inherent part of measurement. That it's not necessarily you sucking, that it's just inherently measurement is uncertain.
Human error	The idea that deviations from an expected value are due to mistakes by the experimenter.	You know, just carelessness in reading numbers...units and conversion of units...that kind of mistakes.
Instrument precision	Concepts related to the inherent uncertainty of a measurement device or apparatus.	It's really important for them to understand the limitations of their instruments, their measurements, and that they even have limitations.
Mean	The concept of the mean or average, and what that quantity represents.	The single most important thing for them to understand is when they made the measurement from N experiments, what is the interpretation of the mean value?
Normal distribution, in particular	The normal or Gaussian distribution as a model for error.	I think of some normal distribution and some Gaussian equation, Gaussian graph, that shows a central value and some spread.
Propagation of error	The idea that the uncertainty of input quantities affects the uncertainty of resulting calculated quantities, and the sensitivity of those calculated uncertainties.	My course teaches the concept that "input uncertainties influence output uncertainties" well.
Relative error	The idea of quantifying error based on the percentage difference between two values, including from a predicted or known value.	So a lot of students get stuck not knowing what a percent or a fractional uncertainty is versus an absolute uncertainty.
Significant figures	The idea that uncertainty can be communicated based on the number of digits in a quantity.	We did discuss a lot about significant figures...that that has that uncertainty built in, right?
Standard deviation	The concept of the standard deviation, and what that quantity represents.	You can use a standard deviation to then say that it gets some uncertainty in the spread...you would expect that it would be sort of Gaussian.
Standard error	The concept of the standard deviation of the mean (or other sampled statistic), such as calculated by dividing the sample standard deviation by $\sqrt{N-1}$.	We can calculate the mean of that distribution [of means], and...that gives us a measure of how much deviation one particular sample mean likely deviates from the true value.
Systematic uncertainty	The idea that error can arise from particular sources, and that such error does not reduce through averaging multiple trials.	We just do some discussion of the systematic uncertainty influencing accuracy, and statistical uncertainty influencing precision, and identifying what could be possible systematic errors.

TABLE V. The set of emergent codes pertaining to measurement uncertainty practices. The first column is the name of the code, the second is a description of that code, and the third is an example excerpt from the interviews.

Code	Description	Sample quote
Calculating error bars	Determining the size of error bars and plotting them on graphs.	I have them calculate some like least squares and some error bars and things for their readings.
Calculating means and standard deviations	Carrying out the calculation of a mean and/or standard deviation.	They have a homework assignment to calculate [the] standard deviation of something.
Interpreting data	Checking if results make sense in general, and if they match reasonable predictions.	We want people to make measurements and then actually think about whether they make sense, right? And not just blindly throw them into spreadsheets and start calculating stuff.
Explaining choices	Deciding what to report, and what is relevant to communicate in a result.	...clarity of thought and writing so that they can explain what they did and...what measurement [they] took, and that gives [them] help with estimating uncertainty.
Determining instrument precision	Deciding and/or estimating the inherent precision of a measurement device or apparatus	We have some printed dials on the page...and they need to tell us the measurement and the uncertainty based on that thermometer or pressure gauge or whatever else.
Estimating error by eye	Deciding on the error of a measurement, such as the spread of a data set, by estimating non-quantitatively (i.e., "by eye").	Then we look at the real data and there's...all these jiggles. And what jiggles are significant, what jiggles are insignificant, how can you tell?
Fitting a function to data	Performing a fit to data, usually using a computer.	I would want them to...graph something in Excel and to do a fit in Excel and get a best fit value with uncertainties.
Formatting results correctly	Reporting results (individual quantities or plots) with appropriate significant figures, and/or with corresponding uncertainties.	I see most often students just kind of neglecting [measurement uncertainty]. So [they] turn in lab reports with averages, but not standard errors.
Identifying outliers	Determining which individual trials in a data set are outliers.	How, for example, if one value is very different, whether to throw it out or repeat it or not.
Identifying sources of uncertainty	Identifying what affects the level of uncertainty of a result, often in the context of experimental design or revision.	I ask...which input variable contributed the most to your final uncertainty? And then I say, which quantity should you measure more precisely to improve this?
Making numerical comparisons	Comparing measured quantities to each other, or to a quantitative prediction, to determine if they agree.	We have two sets of data values. They have to get the average and standard deviation for both sets of values, and then tell us if those measurements overlap.
Propagating uncertainties using brute force	Determining the uncertainty range of a calculated quantity by "plugging in" extreme values of input quantities.	We don't use calculus in the introduction. We do the maximum minimum method. You have to calculate the maximum value based upon the uncertainty of the measurements, and the minimum value, and then propagate the uncertainty through the calculations.
Propagating uncertainties using formulas	Determining the uncertainty range of a calculated quantity using generalized formulas, e.g., from differential calculus.	You have two measurements with uncertainties. What is the uncertainty if we multiply them, if we add them, if we square them?
Using software	Using a computer program, e.g., a spreadsheet program, for recording, analyzing, and/or plotting data.	The nitty gritty boring statistical calculations. In fact I don't really care much if they can do it by hand. If they can get the software to do it...I'm fine with that.
Using terminology correctly	Employing technical phrases correctly when speaking or writing about an experiment.	...a well-written analysis that uses the terminology and the concepts correctly, it's not just working out a problem set and getting the right answer. They talk the talk.

first pass. 10% of the 288 *concepts* excerpts, and 10% of the 245 *practices* excerpts, were used to create two preliminary emergent code books, one describing *concepts* and one describing *practices*. Then, two researchers separately applied these code books to another 10% of the *concepts* excerpts and another 10% of the *practices* excerpts. We calculated the Cohen's kappa statistic [63], a measure of interrater reliability, from these code assignments of the two raters. These calculations yielded an average kappa value of 0.96 for the *concept* codes, and 0.97 for the *practice* codes, both indicating almost perfect agreement [64], which we deemed satisfactory to proceed. Through

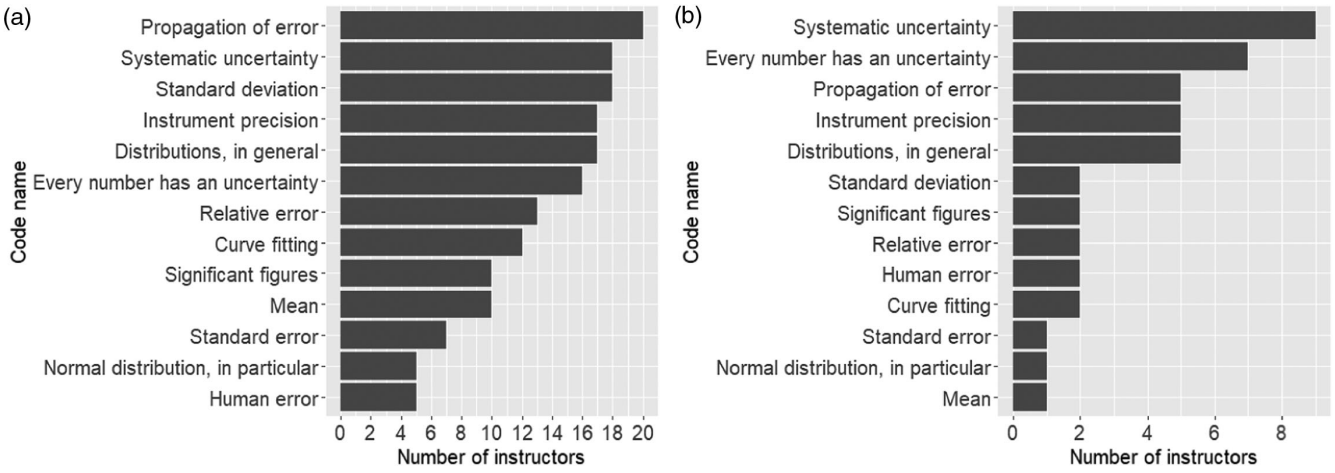

FIG. 2. The number of interviews in which instructors mentioned *concepts* (a), and only those concepts they identified as *important* (b), broken down by the emergent codes in Table IV. As a reminder, we interviewed 22 instructors ($N_{\text{tot}} = 22$).

this process, the wording of each code was clarified to create the finalized *concept* and *practice* code books. These codes are listed in Sec. IV, Tables IV and V, below.

Separately, we used the descriptions of lab *activities* to create a third emergent code book. It was straightforward to identify similar *activities* excerpts, and thus, we deemed it acceptable to create these emergent subcodes all at once, forgoing a formal check of interrater reliability.

3. Second pass

We then used the subcodes in our emergent code books to code the rest of the *concepts*, *practices*, and *activities* excerpts identified in the first pass. The other top-level codes were assigned to excerpts describing both concepts and practices. Therefore, we combined the two emergent books, the concepts subcodes and the practices subcodes, and used this combined code book to analyze the excerpts contained in the remaining top-level codes.

After the entire dataset was analyzed in these two passes, we counted the number of instructors who received each of the emergent subcodes. These counts represent the results of our analysis, with the unit of analysis being each interview in its entirety. We avoid counting the number of times each code was assigned within an interview, as the semistructured nature of the interviews resulted in varied emphasis on particular questions. This variation precluded meaningful interpretation of such finer-grained counting.

IV. RESULTS AND DISCUSSION

The number of interviews that were assigned each *a priori* code related to particular measurement uncertainty topics, and those related to how measurement uncertainty is taught, is shown in Table III. With the exception of normal vs Poisson distributions, which is likely a more advanced topic, each topic was well represented. Likewise, each way measurement uncertainty is taught had substantial representation.

Emergent code books resulted from the second pass analysis of excerpts from the interviews. One represents concepts related to measurement uncertainty, while another relates to practices involved in measurement uncertainty. The concept codes are shown in Table IV, along with a description and a short example from the interview transcripts. These concepts represent an answer to RQ1, the concepts that are relevant for understanding measurement uncertainty according to physics lab instructors. Likewise, the practice codes and corresponding descriptions and examples are shown in Table V. These practices represent an answer to RQ2, the practices that students engage with related to measurement uncertainty according to physics lab instructors.

We now present the number of interviews in which the codes in Tables IV and V were applied, referred to here as counts. We present counts on sets of excerpts that were coded with each *a priori* code from the first pass.

Counts for the *concepts* that instructors mentioned, and only the ones that they identified as *important*, are shown in Fig. 2. The commonly mentioned concepts provide further insight into RQ1, regarding the concepts that are relevant for understanding measurement uncertainty. The most mentioned concepts include "propagation of error," "systematic uncertainty," and "instrument precision." The prevalence of these codes is notable because they all represent concepts beyond the basic mathematical ideas of mean and standard deviation, which often come to mind when thinking about measurement uncertainty at the introductory level. While "standard deviation" is also among the most prevalent codes, the commonality of the other aforementioned concepts shows that instructors have a much broader scope in mind when it comes to measurement uncertainty in their lab courses. This broader scope is further evidenced in the most prominent *important*

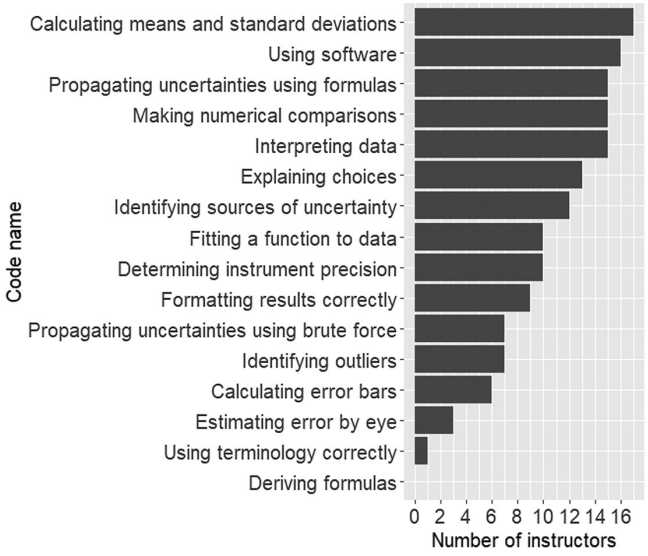

FIG. 3. The number of interviews ($N_{\text{tot}} = 22$) in which instructors mentioned *practices* broken down by the emergent codes in Table V.

concepts, which include "every number has an uncertainty" in addition to "systematic uncertainty".

Counts for the *practices* that instructors mentioned are shown in Fig. 3, and provide further insight into RQ2 as the practices students engage with related to measurement uncertainty. The most common practices, "calculating means and standard deviations" and "using software," represent essential practical skills in modern uncertainty analysis. Additionally, there are other codes that were almost just as prevalent: "interpreting data," "explaining choices," "making comparisons," and "identifying sources of uncertainty," which go beyond these basic practical skills. These codes represent subtasks of the EMF discussed in Sec. II A. The prominence of these modeling elements in introductory lab practices supports a scaffolded learning flow across physics undergraduate lab curricula, and shows how measurement uncertainty acts as the foundation for further lab learning. Lastly, we note that while "propagating uncertainties using formulas" was more common than "propagating uncertainties using brute force," the latter was still mentioned by about a third of the instructors we interviewed. Therefore, the brute force approach should not be discounted when considering how propagation of uncertainty is treated in introductory physics labs.

The remaining *a priori* codes address RQ3, regarding the concepts and practices that are emphasized in introductory physics labs. Counts for topics students already know, separated by whether, from the perspective of each instructor, these topics supported or hindered learning, are shown in Fig. 4. Counts for topics students find difficult are shown in Fig. 5. Counts for topics that they chose not to cover or emphasize are shown in Fig. 6.

Considered together, these results suggest the ways in which concepts and practices come together in introductory physics labs. Students tend to start these courses already knowing about means and standard deviations, both conceptually and as practices, and this knowledge supports their learning. As such, instructors do not talk about the concept of means and standard deviations as much as other concepts, but, as practices, they are still quite prevalent. However, some students are also familiar with the concept of relative error, which can interfere with the measurement uncertainty learning goals of physics instructors. Accordingly, instructors do not emphasize relative error, means, and standard deviations, instead emphasizing propagation of error and systematic uncertainty. Still others avoid these topics entirely, often due to practical limitations and the difficulty of the topic for their students. This duality presents a challenge for assessment development. Our results suggest that propagation of error and systematic uncertainty are simultaneously valuable and difficult

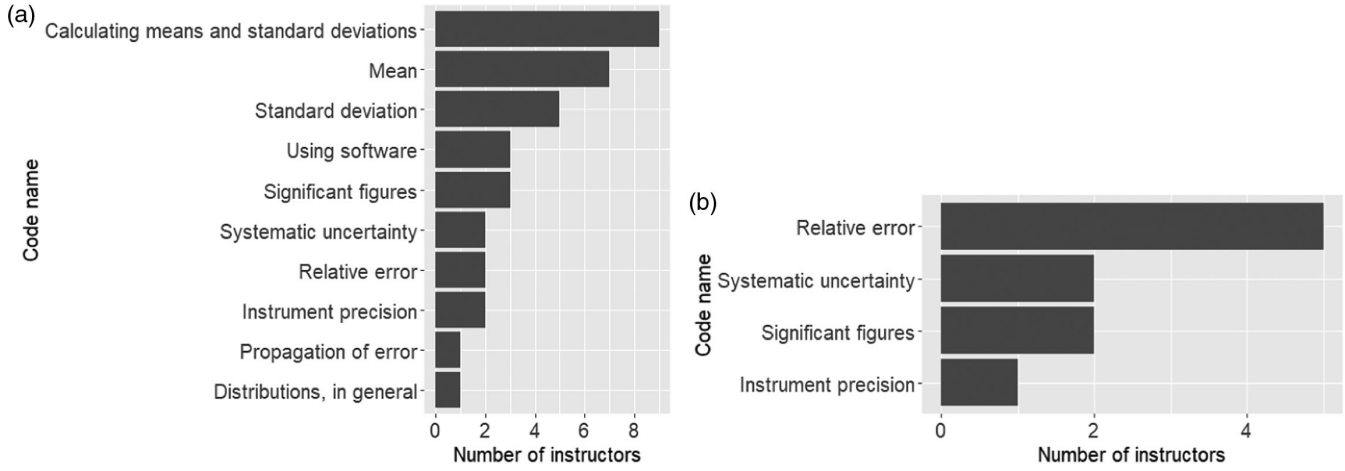

FIG. 4. The number of interviews ($N_{\text{tot}} = 22$) in which instructors mentioned topics students already know, separated by whether these topics supported (a) or hindered (b) learning, broken down by the emergent codes in Tables IV and V.

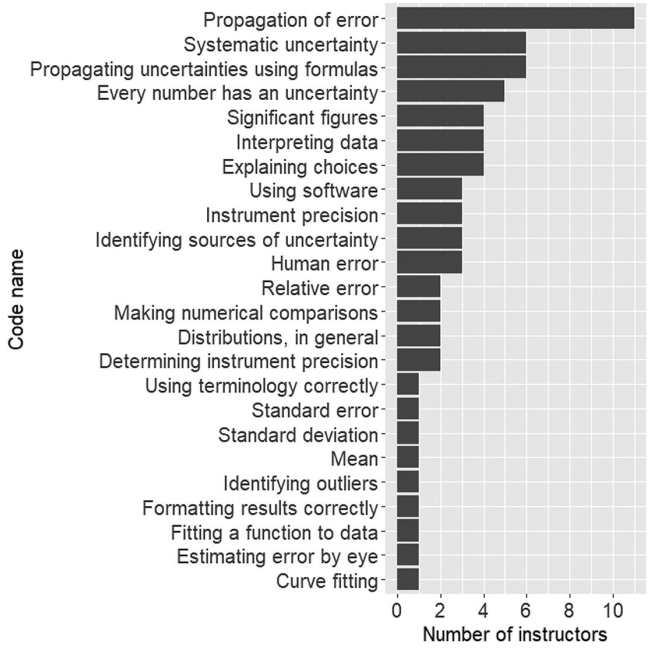

FIG. 5. The number of interviews ($N_{\text{tot}} = 22$) in which instructors mentioned topics students find difficult, broken down by the emergent codes in Tables IV and V.

learning outcomes for introductory labs. Therefore, our results seem to call for two contradictory approaches: both centralizing and omitting these topics from an assessment. The implications of this outcome are discussed further in Sec. V.

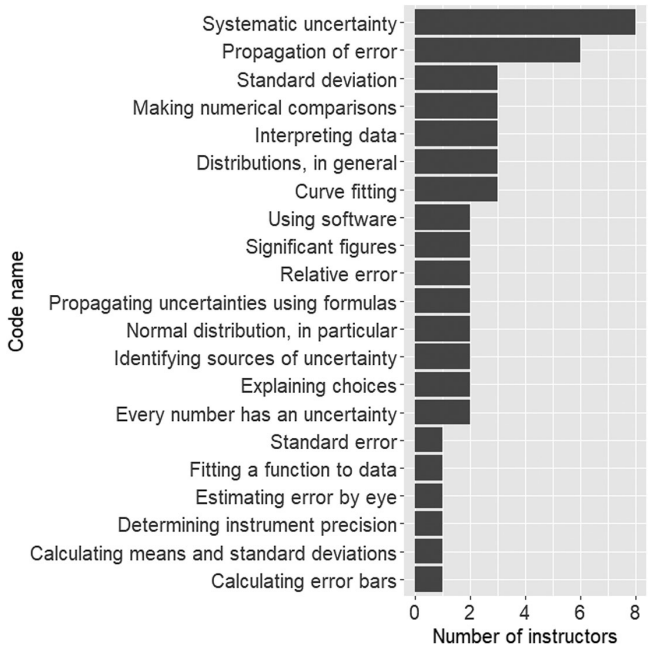

FIG. 6. The number of interviews ($N_{\text{tot}} = 22$) in which instructors mentioned topics that they chose not to cover or emphasize, broken down by the emergent codes in Tables IV and V.

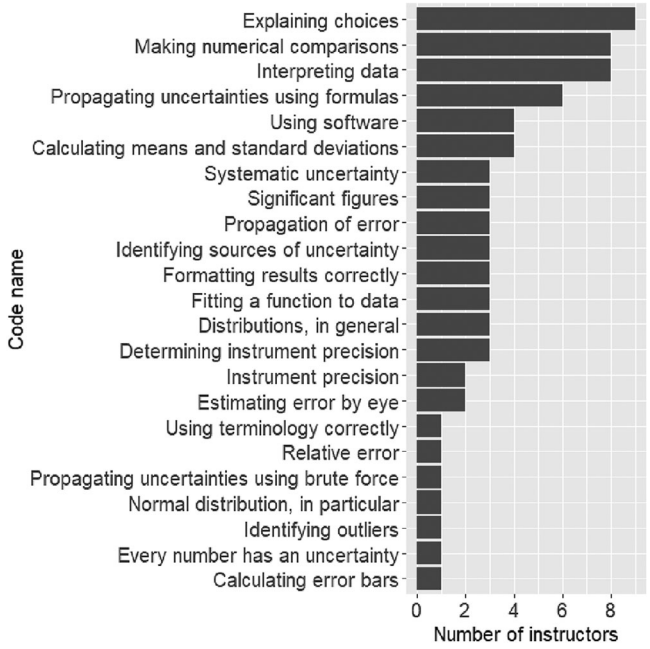

FIG. 7. The number of interviews ($N_{\text{tot}} = 22$) in which instructors mentioned ways of *assessing* measurement uncertainty, broken down by the emergent codes in Tables IV and V.

Counts for how instructors go about *assessing* understanding of measurement uncertainty are shown in Fig. 7. The prevalent topics here are perhaps the most directly applicable to assessment development, as an assessment that mirrors instructors' own approaches to measuring learning is one they would recognize as valid. Here, some of the same previously identified modeling-related practices are prominent, in particular "interpreting data," "Explaining choices," and "Making comparisons." Perhaps these elements of modeling are the focus of assessment because they tie together the more basic concepts and practices of

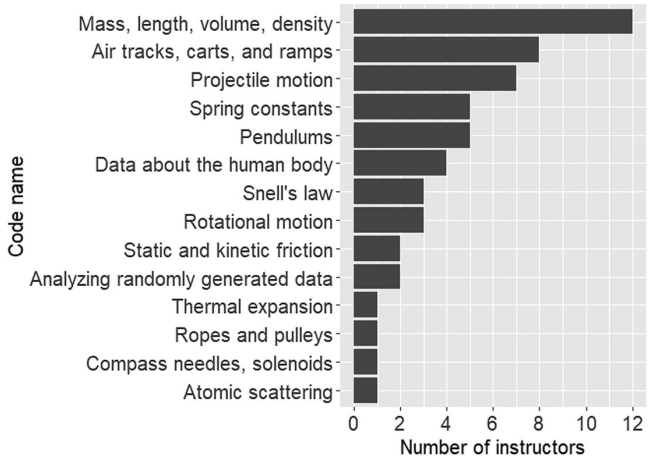

FIG. 8. The number of interviews ($N_{\text{tot}} = 22$) in which instructors mentioned various types of lab *activities* involving measurement uncertainty.

measurement uncertainty, and represent readiness to move on to modeling as a whole in later courses.

Lastly, counts of lab *activities* that involve measurement uncertainty are shown in Fig. 8. The activities that are prevalent in physics lab classrooms will provide context for developing assessment arguments and items, ensuring that the assessment has practical meaning and relevance [26]. These activities further address the contexts mentioned in RQ3.

A. Limitations

Before discussing the implications of these results for assessment development, we note some limitations inherent to the work presented here. While our recruitment process aimed to garner a sample population of instructors from a diverse range of institutions, with only 22 instructors interviewed, we were unable to create a truly representative sample of the diversity of introductory physics lab courses in the US. Thus, while we present counts as an indication of prevalence, these findings should not be taken to truly represent all of introductory physics lab instruction. Instead, we aim for these results to suggest more general themes that are common across different instructional contexts. While they cannot capture the entirety of physics labs, they do provide a view of the range of topics that are relevant to measurement uncertainty in introductory experimental physics.

Lastly, while our sampling aimed to represent a diversity of institutions, the individual instructors who responded to our solicitations were self-selected. Therefore, we cannot guarantee these instructors to represent the typical introductory physics lab instructor, even in their respective institutional contexts. We suspect that the instructors who responded to our outreach would tend to be the ones most invested in their teaching and pedagogy, as such investment is likely the primary motivation for them to take the time to participate in an interview. Nonetheless, similar selection pressures will be at play for the initial adoption of our future assessment, so in that sense, these instructors are representative of a target population for this assessment development project. Moreover, the perspectives of thoughtful and dedicated instructors are perhaps most valuable at this stage of assessment development and have the potential to resonate just as strongly with the population of lab instructors overall.

V. IMPLICATIONS FOR ASSESSMENT DEVELOPMENT

A. Domain analysis

The most prominent emergent codes referenced by instructors will directly inform the creation of SPRUCE by answering three central questions in domain analysis in ECD. In the domain of measurement uncertainty in introductory physics lab courses, we aim to identify (a) what knowledge is constructed, (b) how knowledge is used, and (c) what evidence of student knowledge and proficiency lab instructors will accept.

Point (a) directly pertains to RQ1 and the most prevalent concepts mentioned by instructors, as shown in Fig. 2. Measurement uncertainty knowledge is constructed by understanding not only the statistical concepts of standard deviation and distributions, but also the ideas of sensitivity to variation (under propagation of error), systematic uncertainty, and fundamental philosophies around our ability to measure the true value of a quantity. It will be necessary for SPRUCE to capture this conceptual breadth.

Point (b) directly pertains to RQ2 and the most prevalent practices mentioned by instructors, as shown in Fig. 3. Here, measurement uncertainty is used not just to calculate uncertainty values or error bars, but also to engage in the elements of modeling that come together to interpret and compare data, explain choices, and identify sources of uncertainty. For SPRUCE to be useful in the broader context of undergraduate physics learning, it must contextualize its content in such modeling practices.

Point (c) directly pertains to RQ3 and the concepts and practices identified by instructors in Figs. 4–6, and in particular Fig. 7. The topics that some instructors emphasize overlap with the prevalent concepts and practices they mentioned overall, namely, propagation of error and systematic uncertainty. However, other instructors avoid these topics entirely. As the domain is not monolithic, this finding calls for flexibility in the implementation of SPRUCE. Perhaps certain parts can be optionally omitted from the survey or the analysis, at the discretion of those using the survey, to align with what a particular instructor or researcher feels is relevant [65]. Lastly, when it comes to the ways that instructors assess measurement uncertainty, elements of modeling are again prevalent, underscoring the need for SPRUCE to focus on these aspects.

B. Next steps

Following the overall framework of ECD, we will next create descriptions of coherent and realistic situations that tie together the common concepts, practices, and focuses of physics lab instructors using contexts from their classrooms. This is known as domain modeling, and aims to express the central assessment argument in narrative form [26]. In creating these narratives, we will also bring our own expertise to bear as professional physicists and physics educators.

For example, a narrative could aim to combine the concept of systematic uncertainty with the practices of interpreting and explaining data and identifying sources of uncertainty, in the context of a basic measurement of volume:

> *A measurement of the volume of an irregularly-shaped object is needed to calculate its density in order to determine what material it is made of. These volume*

measurements can be done in several ways, such as by measuring length dimensions and multiplying them to calculate the volume, or by immersing the object in water to measure the displacement. Each of these ways could introduce some systematic sources of error. Students will discuss which measurement approach is best, and how the resulting value for the volume should be presented to those doing the density calculation.

The choice of method requires identifying sources of uncertainty. Different methods introduce different types of systematic effects, such as under- or overmeasuring the volume based on whether its features are concave or convex, or perhaps through the influence of trapped air or evaporation. Interpreting and explaining these effects is necessary to fully present the resulting data to those doing the subsequent calculation.

We will create more narratives such as this that tie together the common concepts, practices, and focuses presented in this work. These narratives will then act as the starting point to developing survey items for SPRUCE.

VI. CONCLUSION

We described here the first step in creating SPRUCE, an RBAI for measuring learning related to measurement uncertainty in introductory physics lab courses. This step, domain analysis, comes from ECD, a framework for developing RBAIs. It calls for identifying what knowledge is constructed and used, and what evidence will be accepted. To accomplish that call, we centered our work around three research questions that ask about concepts and practices related to measurement uncertainty, and which of these concepts and practices are emphasized in introductory physics labs. To address these questions, we interviewed 22 introductory physics lab instructors across the US from a diverse range of institutions. Our analysis showed that prevalent concepts went beyond basic statistical ideas like mean and standard deviation, and included propagation of error and systematic uncertainty. We found that measurement uncertainty was prevalent in practices related to the broader process of experimental modeling, as described by the EMF [2]. We found that such elements of modeling were also emphasized by introductory lab instructors in their assessment practices. Furthermore, propagation of error and systematic uncertainty were emphasized by some instructors, but avoided by others, suggesting the need for some flexibility in the content and analysis of SPRUCE. These results will form the basis for the next steps of assessment development, ultimately leading to a valid and robust RBAI for measuring learning related to measurement uncertainty in introductory physics lab courses at the undergraduate level. The result of this process, SPRUCE, will enable deeper and more complete investigations of measurement uncertainty, a foundational aspect of learning in physics labs.

ACKNOWLEDGMENTS

We acknowledge Michael F. J. Fox for assistance in interrater reliability coding. This work was supported by the National Science Foundation (Grants No. DUE-1914840, No. DUE-1913698, and No. PHY-1734006). The interview protocol was created by B. P., with input from all other authors. R. H. did the coordination and contacting of instructors, and conducted the interviews, with support from B. P. and H. J. L. B. P. checked the transcripts and analyzed the interviews (two interviews were transcribed by R. H.), with the help of another researcher for interrater reliability. B. P., R. H., M. D. C., and H. J. L. contributed to analysis discussions. B. P. wrote the bulk of this manuscript, and M. D. C. wrote a subsection. All authors contributed to editing of the manuscript.

[1] AAPT Committee on Laboratories, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (American Association of Physics Teachers, College Park, MD, 2014).

[2] D. R. Dounas-Frazer and H. J. Lewandowski, The modelling framework for experimental physics: Description, development, and applications, Eur. J. Phys. **39**, 064005 (2018).

[3] N. G. Holmes, C. E. Wieman, and D. A. Bonn, Teaching critical thinking, Proc. Natl. Acad. Sci. U.S.A. **112**, 11199 (2015).

[4] B. R. Wilcox and H. J. Lewandowski, Students' epistemologies about experimental physics: Validating the Colorado learning attitudes about science survey for experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 010123 (2016).

[5] A. Guillon and M.-G. Séré, The role of epistemological information in open-ended investigative labwork, in *Teaching and Learning in the Science Laboratory* (Kluwer Academic Publishers, Dordrecht, 2002), pp. 121–138.

[6] A. Buffler, F. Lubben, and B. Ibrahim, The relationship between students' views of the nature of science and their views of the nature of scientific measurement, Int. J. Sci. Educ. **31**, 1137 (2009).

[7] Joint Committee for Guides in Metrology, Evaluation of measurement data-Guide to the expression of uncertainty

in measurement, Tech. Rep. (Joint Committee for Guides in Metrology, Sèvres, Paris, 2008).

[8] R. L. Kung, Teaching the concepts of measurement: An example of a concept-based laboratory course, Am. J. Phys. **73**, 771 (2005).

[9] R. Beichner, The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project, in *Research-Based Reform of University Physics* (2007), Vol. 1, https://www.compadre.org/Repository/document/ServeFile.cfm?ID=4517&DocID=183.

[10] E. Etkina and A. V. Heuvelen, Investigative science learning environment—A science process approach to learning physics abstract: Table of contents, in *Research-Based Reform of University Physics* (2007), Vol. 1, https://www.compadre.org/Repository/document/ServeFile.cfm?ID=4988&DocID=239.

[11] N. G. Holmes, Structured quantitative inquiry labs: Developing critical thinking in the introductory physics laboratory, Ph.D. thesis, The University of British Columbia, 2014.

[12] N. G. Holmes and E. M. Smith, Operationalizing the AAPT learning goals for the lab, Phys. Teach. **57**, 296 (2019).

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

[14] R. Lippmann Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, NorDiNa **4**, 40 (2006).

[15] N. G. Holmes and D. A. Bonn, Doing science or doing a lab? Engaging students with scientific reasoning during physics lab experiments, in *Proceedings of the 2014 Physics Education Research Conference, Minneapolis, MN* (AIP, New York, 2014), pp. 185–188.

[16] N. Majiet and S. Allie, Student understanding of measurement and uncertainty: probing the mean, in *Proceedings of the 2018 Physics Education Research Conference*, Washington, DC (AIP, New York, 2019).

[17] M. M. Stein, C. White, G. Passante, and N. G. Holmes, Student interpretations of uncertainty in classical and quantum mechanics experiments, in *Proceedings of the 2020 Physics Education Research Conference, virtual conference*(AIP, New York, 2020).

[18] B. Pollard, R. Hobbs, D. R. Dounas-Frazer, and H. J. Lewandowski, Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT* (AIP, New York, 2020).

[19] 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).

[20] A. Madsen, S. B. McKagan, E. C. Sayre, and C. A. Paul, Resource Letter RBAI-2: Research-based assessment instruments: Beyond physics topics, Am. J. Phys. **87**, 350 (2019).

[21] B. Campbell, F. Lubben, A. Buffler, and S. Allie, Teaching scientific measurement at university: Understanding student's ideas and laboratory curriculum reform, Southern African Association for Research in Mathematics, Science and Technology Education (2005), http://www.phy.uct.ac.za/sites/default/files/image_tool/images/281/people/buffler/physics_education/Monograph%202005.pdf.

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

[23] 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).

[24] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking (PLIC), Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[25] National Research Council, *Adapting to a Changing World—Challenges and Opportunities in Undergraduate Physics Education* (National Academies Press, Washington, DC, 2013).

[26] R. J. Mislevy, G. Haertel, M. Riconscente, D. W. Rutstein, and C. Ziker, *Evidence-centered assessment design, Assessing Model-Based Reasoning Using Evidence-Centered Design* (Springer, New York, 2017), pp. 19–24.

[27] B. R. Wilcox, M. D. Caballero, C. Baily, H. Sadaghiani, S. V. Chasteen, Q. X. Ryan, and S. J. Pollock, Development and uses of upper-division conceptual assessments, Phys. Rev. ST Phys. Educ. Res. **11**, 020115 (2015).

[28] D. R. Dounas-Frazer, L. Ríos, B. Pollard, J. T. Stanley, and H. J. Lewandowski, Characterizing lab instructors' self-reported learning goals to inform development of an experimental modeling skills assessment, Phys. Rev. Phys. Educ. Res. **14**, 020118 (2018).

[29] L. Ríos, B. Pollard, D. R. Dounas-Frazer, and H. J. Lewandowski, Using think-aloud interviews to characterize model-based reasoning in electronics for a laboratory course assessment, Phys. Rev. Phys. Educ. Res. **15**, 010140 (2019).

[30] C. J. Harris, J. S. Krajcik, J. W. Pellegrino, and A. H. DeBarger, Designing knowledge-in-use assessments to promote deeper learning, Educ. Meas. Issues Pract. **38**, 53 (2019).

[31] A. P. Jambuge, K. D. Rainey, B. R. Wilcox, and J. T. Laverty, Assessment feedback: A tool to promote scientific practices in upper-division, in *Proceedings of the 2020 Physics Education Research Conference, virtual conference* (AIP, New York, 2020), p. 234.

[32] K. D. Rainey, A. P. Jambuge, J. T. Laverty, and B. R. Wilcox, Developing coupled, multiple-response assessment items addressing scientific practices, in *Proceedings of the 2020 Physics Education Research Conference, virtual conference* (AIP, New York, 2020), p. 418.

[33] N. S. Stephenson, E. M. Duffy, E. L. Day, K. Padilla, D. G. Herrington, M. M. Cooper, and J. H. Carmel, Development and validation of scientific practices assessment tasks for the general chemistry laboratory, J. Chem. Educ. **97**, 884 (2020).

[34] R. Millar, R. Gott, F. Lubben, and S. Duggan, Children's performance in investigative tasks in science: A framework for considering progression, in *Progression in Learning*, edited by M. Hughes (Multilingual Matters Ltd., Clevedon, UK, 1996), pp. 82–108.

[35] 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).

[36] N. Majiet and S. Allie, Student understanding of measurement and uncertainty: Probing the mean, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).

[37] B. Pollard, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab course on students' understanding of measurement uncertainty, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* (AIP, New York, 2018).

[38] 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, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* (AIP, New York, 2018).

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

[40] E. Etkina, S. Murthy, and X. Zou, Using introductory labs to engage students in experimental design, Am. J. Phys. **74**, 979 (2006).

[41] 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).

[42] K. Kok, B. Priemer, W. Musold, and A. Masnick, Students' conclusions from measurement data: The more decimal places, the better?, Phys. Rev. Phys. Educ. Res. **15**, 010103 (2019).

[43] A. E. Leak, Z. Santos, E. Reiter, B. M. Zwickl, and K. N. Martin, Hidden factors that influence success in the optics workforce, Phys. Rev. Phys. Educ. Res. **14**, 010136 (2018).

[44] R. Serbanescu, Identifying threshold concepts in physics: Too many to count!, Pract. Evidence Scholarship Teach. Learning Higher Educ. **12**, 378 (2017).

[45] D. Harrison and R. Serbanescu, Threshold concepts in physics, Pract. Evidence Scholarship Teach. Learning Higher Educ. **12**, 352 (2017).

[46] A. M. Masnick, D. Klahr, and E. R. Knowles, Data-driven belief revision in children and adults, J. Cognit. Dev. **18**, 87 (2017).

[47] D. Hu and B. M. Zwickl, Examining students' views about validity of experiments: From introductory to Ph.D. students, Phys. Rev. Phys. Educ. Res. **14**, 010121 (2018).

[48] N. G. Holmes and C. E. Wieman, *Assessing Modeling in the Lab: Uncertainty and Measurement* (American Association of Physics Teachers, College Park, MD, 2015), pp. 44–47.

[49] M. Vonk, P. Bohacek, C. Militello, and E. Iverson, Developing model-making and model-breaking skills using direct measurement video-based activities, Phys. Rev. Phys. Educ. Res. **13**, 020106 (2017).

[50] M. M. Hull, A. Jansky, and M. Hopf, Probability-related naïve ideas across physics topics, Studies Sci Educ. **57**, 45 (2020).

[51] R. J. Mislevy, L. S. Steinberg, and R. G. Almond, Focus article: On the structure of educational assessments, Meas. Interdiscip. Res. Perspect. **1**, 3 (2003).

[52] National Research Council, *Developing Assessments for the Next Generation Science Standards* (National Academies Press, Washington, DC, 2014), p. 288.

[53] C. J. Harris, J. S. Krajcik, J. W. Pellegrino, K. W. Mcelhaney, A. H. Debarger, C. Dahsah, D. Damelin, C. M. D'Angelo, L. V. Dibello, B. Gane, and J. Lee, *Constructing Assessment Tasks that Blend Disciplinary Core Ideas, Crosscutting Concepts, and Science Practices for Classroom Formative Applications*, Tech. Rep. (SRI International, Menlo Park, CA, 2016).

[54] R. J. Mislevy, R. G. Almond, and J. F. Lukas, A brief introduction to evidence-centered design, ETS Res. Report Series **2003**, 29 (2003).

[55] S. M. Underwood, L. A. Posey, D. G. Herrington, J. H. Carmel, and M. M. Cooper, Adapting assessment tasks to support three-dimensional learning, J. Chem. Educ. **95**, 207 (2018).

[56] R. L. Stowe and M. M. Cooper, Assessment in chemistry education, Isr. J. Chem. **59**, 598 (2019).

[57] L. Crocker, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, New York, 1986).

[58] R. J. de Ayala, The theory and practice of item response theory, in *Methodology in the Social Sciences* (Guilford Publications, New York, 2013), p. 448.

[59] https://carnegieclassifications.iu.edu/.

[60] https://www.aip.org/.

[61] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.010133 for the interview protocol used when conducting interviews with physics lab instructors.

[62] https://www.rev.com/automated-transcription.

[63] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20**, 37 (1960).

[64] N. J. M. Blackman and J. J. Koval, Interval estimation for Cohen's kappa as a measure of agreement, Stat. Med. **19**, 723 (2000).

[65] K. D. Rainey, M. Vignal, and B. R. Wilcox, Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage, Phys. Rev. Phys. Educ. Res. **16**, 020113 (2020).