Modeling novel physics in virtual reality labs: An affective analysis of student learning

Jared P. Canright[*] and Suzanne White Brahmia[†]
Department of Physics, University of Washington, 3910 15th Avenue NE, Seattle, Washington 98195, USA

(Received 26 May 2023; accepted 27 November 2023; published 28 May 2024)

[This paper is part of the Focused Collection on Instructional labs: Improving traditions and new directions.] We report on a study of the effects of laboratory activities that model fictitious laws of physics in a virtual reality environment on (i) students' epistemology about the role of experimental physics in class and in the world; (ii) students' self-efficacy; and (iii) the quality of student engagement with the lab activities. We create opportunities for students to practice physics as a means of creating and validating new knowledge by simulating real and fictitious physics in virtual reality (VR). This approach seeks to steer students away from a confirmation mindset in labs by eliminating any form of prior or outside models to confirm. We refer to the activities using this approach as Novel Observations in Mixed Reality (NOMR) labs. We examined NOMR's effects in 100-level and 200-level undergraduate courses. Using pre-post measurements, we find that after NOMR labs, students in both populations were more expertlike in their epistemology about experimental physics and held stronger self-efficacy about their abilities to do the kinds of things experimental physicists do. Through the lens of the psychological theory of flow, we found that students engage as productively with NOMR labs as with traditional hands-on labs. This engagement persisted after the novelty of VR in the classroom wore off, suggesting that these effects were due to the pedagogical design rather than the medium of the intervention. We conclude that these NOMR labs offer an approach to physics laboratory instruction that centers the development of students' understanding of and comfort with the authentic practice of science.

DOI: 10.1103/PhysRevPhysEducRes.20.010146

I. INTRODUCTION

This study seeks to characterize aspects of student learning that are both highly valued [1] and challenging to assess. In the context of experimental physics courses and using a virtual reality (VR) environment, students engage in activities with novel force laws that are designed to meet a need for introductory laboratory activities that deepen undergraduate physics students' understanding of the process of generating *new* knowledge in science, and the quantitative scientific scrutiny involved. The objective of this study is to better understand the extent to which exploring novel physics, made possible through the use of immersive technologies, can render students more expertlike in their beliefs (i) about how scientific knowledge is generated and (ii) in their capacity to produce scientific knowledge.

In light of the physics education research community's current understanding that laboratory instruction is not an effective means of teaching conceptual content [2,3], we instead seek to use labs as a place where students engage in the authentic practice of science, equipping them with the tools to understand the world through an empirical lens in alignment with the AAPT's recommendations for undergraduate physics lab instruction [1]. We aim to foster an expertlike understanding of the role and process of experimentation [4,5], build their confidence in their ability to design, perform, and interpret experiments [6,7], and keep them actively engaged through the whole process [8].

Understanding how to provide engaging opportunities for students to develop mathematical models of novel phenomena in a teaching laboratory is a difficult open problem in physics pedagogy [1,2,9]. This kind of divergent, creative activity is fraught with challenges related to student autonomy and safety, opportunities for meaningful contexts [9], and the expertise of instructors to engage in a manner that responds to what is happening within each group. These issues are particularly challenging in large-enrollment courses where labs are commonly taught by inexperienced undergraduate and graduate teaching assistants (TAs).

Our framework for designing activities in which students learn to generate models is the investigative science learning environment (ISLE) approach [5,10–13] by Etkina *et al.* In ISLE, model generation happens during what the authors have named *observational experiments*, where students engage in open-minded exploration with the

[*]jpcan@uw.edu
[†]brahmia@uw.edu

goal of developing a model for an unknown phenomenon. This phase of the ISLE process, itself a simplified but authentic representation of the scientific process, is followed by iteratively testing, revising, refining, and applying the model. In our approach, we alter the language slightly from Etkina to optimize transparency for the students of what they are doing. We refer to the processes of *model generating* and *model testing*, rather than observational and testing, experiments. Note that, in the context of ISLE, the terms "model," "explanation," and "hypothesis" are interchangeable [14].

Model-generating experiments involving novel scenarios are difficult to create, especially in introductory courses. In many cases, experimental physics questions that reasonably could be investigated at the introductory level are well known with easily Googled answers. In the presence of known answers, students tend to hold those in the highest regard, seeking to confirm known answers above anything and everything else, even in the face of contradictory data [15] or explicit instruction to the contrary. Thus, any access to a "right answer" can derail efforts to engage students in authentic model generation. We address this expectation of getting the right answer by putting students into a different universe with *new* physics that builds on Newton's laws and fundamental conservation laws—where neither they, nor their textbook, nor Google, nor their TAs have a ready-made model at hand. The problem of shifting students' mindset toward generating new models becomes trivial when there are no existing models to confirm.

The activities described in this paper have been part of the University of Washington (UW) introductory physics curriculum for over 2 years. In the Novel Observations in Mixed Reality (NOMR) labs [16,17], students explore real and fictitious physical phenomena in an immersive 3D environment. Instructors are struck by the ways that students mature as scientists through these labs, an impression that is not easily quantified. The following is an excerpt from a postcourse survey that is fairly typical at the sophomore level and representative of those most impacted at the introductory level.

> VR labs were fantastic for learning how to effectively approach a physics situation where I didn't already know what would happen. In most experiments I have done in previous courses, I had learned what to expect before I was actually making observations and collecting data, so this course helped me learn a new way to approach experiments.

This study is a step toward characterizing this kind of intellectual growth we observe in many students. The work is situated in efforts across the physics education community to find and adapt affective assessment tools beyond standard course evaluations. Our study seeks to establish whether students' belief in their ability to do physics and their sense of belonging in physics grow along with their understanding of the nature and role of experimental physics in generating new knowledge. We assess the impact of the intervention on students: epistemology about experimental physics; physics self-efficacy; and engagement in the learning process. This study contributes to the ongoing research into assessment of what students take away from effective laboratory instruction [18,19]. Specifically, we focus on the following research questions:

RQ1 What changes are observed in students' epistemology about experimental physics as a result of the NOMR labs?

RQ2 What changes are observed in students' physics self-efficacy in experimental physics as a result of the NOMR labs?

RQ3 To what extent are students productively engaged in the NOMR activities, and how does that engagement compare with the hands-on labs in the same course?

II. BACKGROUND

A. ISLE

We begin by considering what lab activities reflecting the real-world practice of science look like. The ISLE approach to physics education [5,10–13] prioritizes epistemologically authentic investigation of physics as a means to develop students' scientific abilities [7] and habits of mind. Teaching students to think like expert physicists takes priority over covering conceptual content.

Three types of experiments form the core of ISLE instructional activities, related to each other by the ISLE process (Fig. 1):

Model-generating experiments: Labeled in the diagram as observational experiments, students engage in open-minded exploration of a previously unknown physical phenomenon. They make note of patterns in

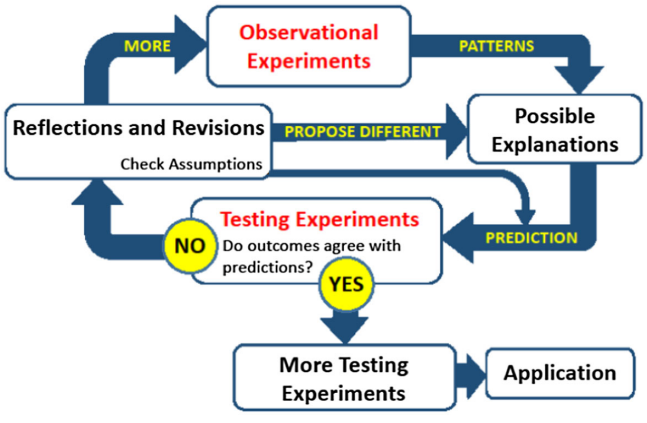


FIG. 1. The ISLE process is a simplified representation of the real-world practice of science, iteratively generating, testing, rejecting, and refining models to empirically create and validate knowledge.

the phenomenon's behavior and devise explanations for those patterns. These patterns become mathematical models and mechanistic explanations of the phenomenon. The models created in model-generating experiments form the basis of students' knowledge of the phenomenon.

Testing experiments: A model from a prior model-generating experiment is tested. Students design an experiment with well-determined independent, dependent, and controlled variables to test a prediction about the outcome of an experiment that follows from the model in question. They run the experiment, collect and analyze their data, and judge whether the outcome is consistent with the prediction. If so, they have supported, or failed to reject, the model. If not, the model is rejected.

Application experiments: Students apply tested models to determine the value of unknown physical quantities or solve practical problems.

In the ISLE approach, content and process are considered to be inextricably paired; these three types of activities are how students uncover new physics content. Students encounter physical phenomena for the first time through hands-on experimentation, and only after identifying patterns and developing their own explanations do they read about the phenomena in their text.

ISLE-approach labs consist primarily of questions to guide students' thoughts rather than dictate them. In this way, students learn to ask and answer questions in the way that a scientist would. This process is guided and refined by scientific abilities rubrics [7] used to assess and give feedback on their work.

B. Affective measures

This study employs three different research-validated surveys to explore different aspects of students' engagement and learning. Table I gives the name, abbreviation, target metrics, format, and administration schedule for each survey.

1. Epistemology about experimental physics

The lab epistemology survey (LES) originally developed by Hu and Zwickl [20,21] is used in this study as a measure of students' epistemology about the role of experimental physics in class and in the world. The LES was developed as an instrument to characterize the beliefs held by physics students at introductory undergraduate, upper-division undergraduate, and graduate levels about experimentation, models, and their roles in the scientific process. This study uses the LES pre-post to assess changes in students' epistemology about experimental physics before and after students complete the NOMR labs. Measuring changes in students' epistemology gives us a window into whether they are adopting the beliefs, attitudes, and mindset about experimental physics characteristic of expert physicists.

The LES is composed of six open-ended questions, accompanied by a codebook used to identify themes in student responses in a consistent and reproducible way. We focus on the first two questions:

LES1 In your opinion, why are experiments a common part of physics classes? Provide examples or any evidence to support your answer.

LES2 In your opinion, why do scientists do experiments for their research? Provide examples or any evidence to support your answer.

Novice responses to these LES items exhibit an almost singular focus on the idea that experiments in instructional labs exist to supplement conceptual learning or test theories (using a layperson's understanding of the term "theory"). The term "theory" is somewhat vague here: It has a specific definition in the context of physics but is used in a lay sense by students, often translating to "anything that is not an experiment" or "what we know from the textbook or lab manual." Expertlike responses more frequently acknowledge the role of in-class experimentation in the development of scientific abilities and as a means to better understand the scientific process. With regard to experiments in scientific research, novices tend to focus on the notion that experiments exist to test theories. Experts more

TABLE I. The characteristics and administration schedule of each survey used in this study are summarized.

Survey		Target metrics	Format	Schedule
Lab epistemology survey	LES	Student attitudes and beliefs about the nature and role of experimentation in physics	Five open-ended short-answer questions	Presurvey at the beginning of the term; 100-level postsurvey after NOMR activities concluded; 200-level postsurvey after course final
Physics identity survey	PhIS	Degree of students' self-identification as a physicist, interest in physics, and belief in their ability to practice and succeed at physics	Six 6-point Likert items (self-perception and interest); Five 7-point Likert items (self-efficacy)	
Flow survey	FS	Degree and nature of students' engagement with the week's lab activity	Seven 7-point Likert items	Weekly at the conclusion of each lab activity

frequently cite the creation of new models and the iterative nature of experimental model development as purposes of experiments in research.

The original LES [20] included the codes *theory testing* ("The purpose of doing physics experiments is to prove a theory or test a hypothesis.") and *theory development* ("Experiments inspire the development or improvement of theories."). Due to the vague nature of the term "theory," it is unclear how Hu and Zwickl drew a distinction between theory and hypothesis as used by students. To use the term theory in our analysis, we would need to establish our own definition, at risk of misrepresenting students' responses in a replication study.

In alignment with ISLE, we remove references to the term theory in favor of the term model. A model is a foundational concept in ISLE, used heavily throughout both populations' lab activities. The modified codes we use in our analysis are *Model* testing ("The purpose of doing physics experiments is to prove, support, or test a model.") and *Model* development ("Experiments inspire the development or improvement of models.").

The distinction between *Model development* and *Discovery* (original definition: "Experiments help investigate unknowns.") is subtle and not one we are interested in probing; rather, we are more interested in understanding whether students' responses reflect any acknowledgment at all of the steps of the ISLE process aside from model testing. Instead, we collapse the two codes into a single *Discoverment* code: "Experiments contribute to some aspect of the iterative and generative nature of the scientific process aside from testing an existing model."

The final code list is given with definitions and examples in Table II.

TABLE II. LES items, the codes associated with each, and an example response tagged with each code are shown. The example responses for each code were selected such that each example was assigned only to the associated code. Many responses in the data were assigned more than one code. *Scientific abilities* and *Supplemental learning* are drawn from Hu and Zwickl's original codebook [20]. *Model testing* is equivalent to the original codebook's *Theory testing*. We added the *Modeling* code in response to the recurring presence of its ideas in our dataset and its relevance to our research questions. The original codebook's *Discovery* and *Theory development* codes were merged to create *Discoverment*.

Item	Code	Definition	Example student response
LES1: Why are experiments a common part of physics classes?	*Modeling*	Experiments in class let students develop their own models for phenomena, discover things on their own, and/or develop their own ideas.	Because it helps show the process of developing a model, rather than just taking it as fact and using it to solve problems. By studying "mystery particles" in lab, we had to experiment and develop our own observations.
	Scientific abilities	Experiments help cultivate students' scientific abilities, such as experimental design, data collection, and data analysis skills.	Experiments provide a way to provide reasoning skills as applied to physics, of which experiments *[sic]* reasoning is needed to have problem [solving] skills not only in the course but in other aspects of life.
	Model testing	The purpose of doing physics experiments is to prove, support, or test a model.	Experiments are necessary to test theories. Theories cannot be made into laws without testing.
	Supplemental learning	Experiments provide supplemental learning experiences for concepts and theories.	Experiments are a common part of physics courses because they help you understand the concepts we are learning.
LES2: Why do scientists do experiments for their research?	*Discoverment*	Experiments contribute to some aspect of the iterative and generative nature of the scientific process aside from testing an existing model.	Scientists in the real world are consistently working to provide new findings that deepen our understanding of the world. [There are] plenty of examples in the past, including Newton's laws of motion and evolutionary theory.
	Model testing	The purpose of doing physics experiments is to prove, support, or test a model.	You cannot confirm a hypothesis without performing experiments. Without gather *[sic]* data you cannot decide if something is true or not

2. Physics self-efficacy

In the context of the physics classroom, self-efficacy refers to students' belief in their ability to practice and succeed at physics. Developing students' self-efficacy is a primary goal for our laboratory instruction, as we want students to walk away from the course with confidence in their ability to design, perform, and interpret experiments [6,7].

This study employs the physics identity survey (PhIS), which we adapted from a science identity survey administered to middle school biology students to evaluate shifts arising from their participation in an immersive virtual lab [22]. The original survey was developed through the lens of Hazari's science identity framework [23].

The PhIS is divided into two sets of items probing (i) self-efficacy and (ii) physics identity and interest. We focus on the self-efficacy items, listed below, each on the scale [1: Not at all confident—7: Completely confident]:

PhIS1 How confident are you that you can design an experiment to answer a scientific question in physics?

PhIS2 How confident are you that you can look at the data that you collect and characterize its patterns mathematically?

PhIS3 How confident are you that you can understand the kinds of problems that experimental physicists would investigate?

PhIS4 How confident are you that you could contribute to a team of physicists investigating an experimental physics problem?

PhIS5 How confident are you that you can defend your data analysis to a team of expert physicists?

These items were adapted by swapping out learning goals of the ecosystems biology course of the original study for learning goals of the lab courses of concern in this study, e.g., designing an experiment to answer a scientific question in physics and mathematically characterizing patterns observed in data.

We validated the Likert scale items comprising the PhIS in accordance with Adams and Weiman's recommendations for the development of formative assessment instruments [24]. We conducted think-aloud interviews with eight 100-level physics students. Participants were asked to rate each Likert-scale item and explain their choice as they did so. Participants were recruited through an announcement over the course's web page and incentivized to participate with $20 gift cards. The interviews were conducted online, audio recorded, and transcribed by Otter.ai and subsequently handcorrected.

The interview transcripts were examined to assess the alignment of students' reasoning for the responses they chose with the construct each item was meant to assess. Students' understanding of each item reflected our expectations and their reasoning for each choice revealed nothing unexpected. The validation interview results did not lead to any modification of the PhIS.

3. Flow as a measure of engagement

We use the psychological theory of flow pioneered by Csíkszentmihályi [25] as a lens through which to examine students' engagement with class activities. Known colloquially as being "in the zone," flow is described as a state in which one is completely absorbed in an activity for its own sake, where one action leads smoothly into the next, and one's sense of time becomes distorted. A balance between the person's self-perceived skillfulness at and the challenge posed by an activity is instrumental to achieving a flow state; great challenge must be met with commensurate belief in one's own skill. It is in this state that the most effective learning happens [26].

Csíkszentmihályi identified seven conditions for a person to achieve flow:

1. They know what to do (a clear goal).
2. They know how to do it.
3. They are receiving clear and immediate feedback to know how well they are doing.
4. They know where to go (if navigation is involved).
5. They see what they are doing as challenging.
6. They are confident in their ability to complete the task.
7. Their environment is free of distractions.

As the flow model is fundamentally one of engagement with an activity, it has utility as a measurement of students' engagement with the learning process [26–28]. Active engagement is key to learning [8]; conversely, even well-designed instructional activities with epistemologically authentic inquiry as in ISLE cannot reach students who are not engaged in the learning process.

A subset of the flow conditions comprises an effective basis for maintaining task involvement: A learner needs feedback, confidence in their ability to complete the task, and an environment free of mental distractions. To stick with a task to completion, it is critical for the student's self-efficacy to be great enough that they believe they can do so [6]. To support this belief, they need clear feedback to know how well they are doing and what the next steps are.

Rebello and Zollman note [27] that the zone of proximal development [29], the optimal adaptability corridor [30], and flow are all representations of a balance between a learner's skill and challenge. To express the optimal adaptability corridor's dimensions in terms of flow, horizontal transfer (efficiency) maps to skill, and vertical transfer (innovation) maps to challenge. Flow comes in as a means to tie this balance to other affective elements of the student experience, unifying a number of affective constructs in educational psychology under one quantifiable umbrella.

Massimini and Carli's efforts [31,32] to develop a quantitative instrument to measure flow led to the eight-channel flow model we use in this study. One begins by constructing a mental state diagram, hereafter, referred to as a flow plot, with perceived knowledge and skillfulness on

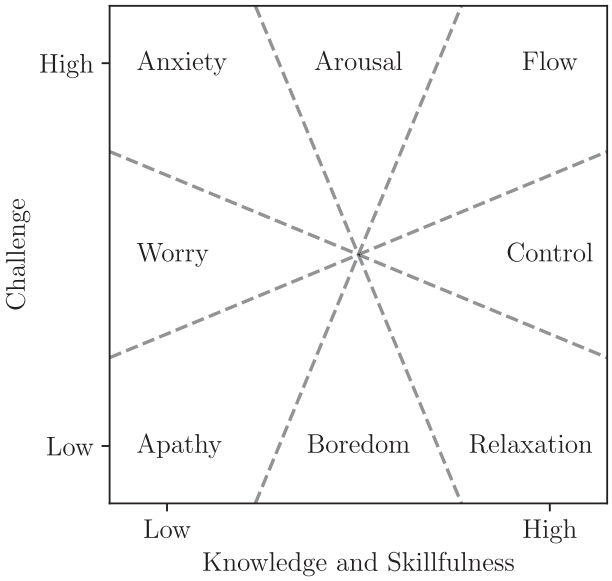


FIG. 2. The eight-channel model of flow.

the horizontal axis, and perceived challenge on the vertical axis. Each flow plot is divided into eight channels (Fig. 2), representing different relative combinations of challenge and skill. The top-right channel, flow, is the most productive, representing a great challenge met with commensurate skill. Flow's neighbor channels control and arousal represent less challenge and less skill than flow, respectively. Flow, control, and arousal are considered productive channels for learning [26]. The relaxation channel represents a surplus of skill and dearth of challenge; its mirror channel anxiety represents an extreme challenge one feels poorly equipped to handle. The least productive states of worry, apathy, and boredom fill out the lower regions of the challenge-skill space, with skill and challenge both insufficient to support productive engagement.

Karelina *et al.* [28] used the eight-channel model to compare students' engagement with content-equivalent ISLE-aligned hands-on and video-based labs; we follow much of their methodology in our study of students' engagement with VR labs.

The flow survey (FS) is drawn from Karelina *et al.*'s adaptation of a subset of items from the psychometrically validated Flow State Scale [33]. The FS uses items from their adaptation with minor wording changes for our experimental context.

It consists of seven 7-point Likert scale items:

F1 To what extent was the instructor's assistance needed? [1: Not at all—7: A lot]

F2 To what extent did you know what to do (goal of the task)? [1: Not at all—7: A lot]

F3 To what extent did you know how to do it? [1: No idea—7: Completely]

F4 To what extent did you know how well you were doing? [1: No idea—7: Completely]

F5 To what extent was the lab challenging? [1: Not at all—7: Extremely]

F6 To what extent did you feel knowledgeable and skillful during the lab? [1: Not at all—7: Extremely]

F7 To what extent was the lab fun and interesting? [1: Not at all—7: Extremely]

We follow Karelina *et al.*'s analysis methods to create flow plots using two items: **F5** as a measure of perceived knowledge and skill on the horizontal axis, and **F6** as a measure of challenge on the vertical axis. Students who give a high score to both items are understood to be in a flow state.

Karelina *et al.*'s study compared average responses along each axis of the flow plot with two-tailed paired t tests to determine differences in students' perceived skill and challenge between video and hands-on treatment groups. These quantitative comparisons were backed by visual comparisons of which channels students tended to fall in each treatment group.

We also make use of **F7** ("...fun and interesting?") to characterize the effect of the novelty of VR on students' experience. Weekly measures of this item allow for comparison across lab activities, e.g., comparing hands-on labs to VR labs.

III. METHODS

A. Structure of NOMR labs

The intervention examined in this study uses VR labs to allow students to experience and analyze physical laws in the context of particle interactions that do not exist in nature or on the Internet. We include the constraint that they are consistent with our universe's physics so that students can rely on their extant physics knowledge when reasoning in the VR space. These fictitious physical laws can be construed as hypothetical mathematical variations of Coulomb's law. A selection of fictitious phenomena is described in more detail in Ref. [16].

The virtual apparatus is designed such that it does not give perfect answers; experimental uncertainty is still very much present even in the simulation. The "right answers" programmed into the simulation are never shared with students or their TAs, such that the only "right" answer is the one that students can make the best case for.

In the virtual lab space, students can access force and distance measurement tools and a supply of particles (modeled as hard spheres) exhibiting the behavior(s) they are investigating. These particles can be moved around the space freely, as well as be fixed in place individually. To facilitate creating static arrangements of particles, physics can be temporarily paused in the entire space. As there is no copy of the lab manual nor any means by which to record data while in the headset, the operator relies on their group's interaction and record keeping. Each group of 3–4

students shares one VR headset, with its display mirrored onto a lab computer.

Multiple instructional practices are in place to combat gendered task division common to inquiry-based physics labs [34,35]. All groups complete a teamwork agreement at the beginning of the term outlining expectations and norms for their interactions in and out of class. The NOMR lab manuals prompt students to take turns using the headset at multiple junctures. Students are encouraged to have each member of the group use the headset to collect data, as a means of obtaining multiple measures from which to determine a central value and uncertainty for each of their measurements.

The NOMR labs described in this study are used in two instructional contexts: in introductory calculus-based physics and in a sophomore-level lab for applied physics majors. The first lab encountered, called Charge and Mint, comprised two activities. Introductory students complete these components as two separate labs (VR1 and VR2; see full lab titles and schedule in Table III), and students in the advanced course complete both components in a single lab session (VR1 + 2).

First, groups design and conduct an experiment to test whether virtual analogs of electrically charged particles follow some rescaled version of Coulomb's law. This lab serves to familiarize students with the VR environment and doubles as an opportunity to teach (or review) data linearization.

Second, they take qualitative and quantitative data to create an empirical model for the interaction between fictitious *minty* particles, which behave according to an unknown force law. That force law is not included here in an effort to keep it out of print; instead, we present a handful of observations students might make and leave the specifics of the model to the reader's imagination:

- Minty particles repel when they are near each other and attract when they are far away. A turnaround point where the force is zero exists at a certain separation between particles.
- If two minty particles are brought as close as possible to one another and released from rest, they appear to undergo oscillatory motion. Considering the full range of distances achieved in this motion, the range of distances for which the force between the particles is repulsive seems to be shorter than the range of distances for which it is attractive.
- Beyond the repulsive region, no matter how far apart two minty particles are moved, they continue to exert upon each other a substantial attractive force that increases with distance.

Charge and Mint are used as a preparatory lab to get students familiar with the virtual learning environment and comfortable with the idea of developing a mathematical model for a completely unknown phenomenon. Once the preparatory lab is complete, students are given a new, more complex phenomenon to explore and model, without any phenomenon-specific scaffolding. This final lab takes two forms: the one-week Exotic Matter Lab for introductory students (lab VR3), and the 3-week Manifold Lab (VR3–VR5) for advanced students.

The Exotic Matter Lab's content is identical to the first week of the Manifold Lab: Every group is assigned a different set of fictitious phenomena (referred to as a *scenario*). This phase allows for differentiated instruction as the TA assigns scenarios at the start of class based on their impressions of each group's strengths and weaknesses and the nature of the challenge each scenario poses. Each scenario contains up to three distinct types of particles, all visually identical on creation, picked from at random whenever the user creates a new particle. The phenomena

TABLE III. This timeline of events shows the curriculum 100-level and 200-level students completed over the course of the quarter and when each survey was administered to each population. All LES and PhIS pre- and postsurveys were administered outside of class time except for the 200-level post-test, which students completed in class after their final presentations. The labels **G**, **T**, **A**, and **C** represent model *generation* experiments, model *testing* experiments, *application* experiments, and *communication*, respectively. Note that T = Testing, G = Generating, A = Application, and C = Communication.

	100-level			200-level		
Wk	Lab	Name	Type	Lab	Name	Type
1	0	Coulomb's law	T	E1	Electron beam pt 1	T
2	VR1	Charge	T	E1	Electron beam pt 2	C
3	VR2	Minty	G	N1	Nuclear decay pt 1	T
4	VR3	Exotic matter	G	N2	Nuclear decay pt 1	C
5	B1	Bulbs pt 1	G	VR1 + 2	Charge + Minty	T/G
6	B2	Bulbs pt 2	T	VR3	Exotic matter	G
7	B3	Capacitors	G	VR4	Manifold proposal	C
8				(No lab)		
9	B4	Unknown resistor	A	VR5	Manifold testing	T
10		(No lab)		VR6	Final presentations	C

underpinning each scenario, and the subject of students' inquiry, are the force laws governing the interactions between the particles. In most cases, a single force law dictates the interaction between each pair of particles, though students may develop different valid interpretations supported by their data. The force laws programmed into NOMR are never shared with students or TAs.

In this model-generating experiment, students are told that they have at most three distinct types of particles in their scenario and given tools to label particles and temporarily remove particles from play. Their goals are to determine how many types of particles they have, develop a procedure for identifying an unknown particle, and come up with a testable empirical model describing some subset of the behaviors they observe. Students write up their findings in a full lab report. For introductory students, this model-generating experiment report marks the end of their foray into VR.

Advanced students working through the Manifold Lab instead submit reports describing their model-generating experiments and resulting models to a classwide repository. During class in the second week, each group selects another group's report describing a model of a scenario they had not yet interacted with themselves. They write and submit a proposal before the third week of lab, describing an experiment to test the other group's model. These experiments are carried out in the third week, and their results presented in an oral talk symposium in the fourth week.

The Manifold Lab is presented to students with a gamelike narrative in which they function as research scientists. They explore different "pocket" universes with novel forms of matter obeying fundamental force laws unknown to our universe. The privilege to conduct the second experiment with better equipment depends on applying (noncompetitively) for grant funds: Before performing the second experiment, students write a single-page "grant proposal" in which they summarize another group's findings, propose an experiment to test their model, and request additional or improved equipment within the VR lab. The instructor serves as an entity equivalent to the NSF and its reviewers: They review students' grant proposals for feasibility and work with each group to revise proposed experiments such that it is likely that each will produce a clear outcome that builds on the prior group's findings. Each group receives a few credits to spend on equipment, e.g., more precise measurement tools, a larger workspace, tools that snap to more convenient configurations, or the ability to automatically pause physics after a set amount of time. Occasionally, a clever idea from a group of students inspires the development of a new tool in NOMR, which is added to the upgrade options going forward.

This design seeks to emulate the experience of working within a professional scientific collaboration: The class as a whole collaborates by sharing data and designing experiments to test and revise each other's models. In doing so, students complete an entire cycle of the ISLE process: One group creates a model through a model-generating experiment, another tests it with a testing experiment, and those results serve to reject, revise, or further substantiate the model.

B. Instructional context

The study activities took place at the University of Washington in Seattle (UW), a large R1 public research university in the Pacific Northwest. Of the population of students enrolled in the courses examined in this study, 65% identify as male and 35% as female (nonbinary gender identities are not reflected in UW records, and we did not solicit this information from students separately). White (41%) and Asian (36%) students make up the majority of the population, followed by students who identify with two or more races (8.7%), Hispanic or Latino,a,e students (7.6%), and Black students (2.8%).

Our data come from two physics courses at UW during Fall 2022. All instruction was held in person except in the event a student could not attend a lab due to illness, in which case their lab partners brought them in via video call, when possible. In both courses, groups of 3–4 students worked together for the entire quarter. Each class's lab curriculum schedule is shown in Table III.

100-level population: NOMR was implemented during calculus-based electromagnetism, the second of a three-quarter introductory physics sequence. 467 students enrolled in Fall 2022 and 380 students consented to participate in the study. We refer to the consenting students as the *100-level population* hereafter. This course consisted largely of engineering (74%) and science (17%) students filling prerequisites for their major. Students met weekly for three 1-h lecture sections, a 1-h tutorial section, and a 2-h lab section.

200-level population: Introduction to experimental physics used NOMR as well. This course enrolled 38 students, mostly applied physics majors (55%) who typically intend to follow an industry-oriented path after graduation, alongside other physics and astronomy majors (21%), prescience majors (13%), computer science majors with physics minors (8%), and one math major. All 38 students consented to participate in the study; we refer to them as the *200-level population* hereafter. Students met weekly for a 1.5-h lecture section and a 3-h lab section. Eight members of the 200-level population had previously seen NOMR labs in the modern 100-level labs; all other 200-level students had taken traditional or online (due to COVID) 100-level labs, without VR.

1. Lab activities

All lab activities in the 100- and 200-level courses are designed in alignment with the ISLE approach. We say "in

alignment" because a full implementation of ISLE requires integration of the ISLE process across all components of a course (lecture, lab, etc.), which is not the case at UW.

Every lab activity can be categorized as a hypothesis generating, hypothesis testing, or application experiment (as in Table III), excluding weeks in the 200-level course dedicated specifically to writing and communication. Students' work is guided and assessed with the ISLE scientific abilities rubrics [7].

The first four weeks of the 100-level labs (labs 0, VR1–VR3) focus on particle interactions. Lab 0 is a qualitative testing experiment on Coulomb's law. Students test the effects of charge and separation on the electric force between a copper sphere and a Teflon rod. This is the simplest lab of the quarter and deliberately so. It is the first lab of the quarter when students are still joining the course and switching between sections. The following three weeks (labs VR1–VR3) are NOMR labs, as described in Sec. III A.

The remaining 4 weeks of labs are traditional hands-on labs exploring circuits:

- **B1**: Students begin their exploration of circuits with a model-generating experiment seeking to develop a model to describe the behavior of battery-bulb circuits in series, parallel, and mixed configurations.
- **B2**: Models generated in the prior week are tested against a mixed-configuration circuit. Students create predictions for the current and voltage through each element of the circuit based on their model from the prior week, build the circuit, collect data, and compare the results with their predictions. Where there is disagreement, students revisit and revise their model.
- **B3**: Capacitors are introduced. Students perform a model-generating experiment to develop a mathematical model for the voltage across a charging capacitor.
- **B4**: Students are given a resistor of unknown value and a model of voltage across a discharging capacitor. Using this model, students perform an application experiment to determine the value of the resistor, with uncertainty, by manipulating the model to give the resistance in terms of the slope of a linearized plot and the capacitance of the capacitor.

The 200-level course opens with the electron beam lab (E1–E2). Each group is given an electron beam apparatus (commonly called an "e/m apparatus") that fires electrons across a user-specified voltage into a helium-filled bulb subject to an approximately uniform magnetic field generated by Helmholtz coils outside the bulb. Students are asked to devise and answer a scientific question with the apparatus. Most often, this ends up being a testing experiment based on students' knowledge of electrons' motion in a magnetic field. Occasionally, it turns into a model-generating experiment if a group does not recall this model.

The subsequent nuclear decay lab (N1-N2) works in a similar fashion: Students are given an apparatus, instructed in its operation, and are set loose to devise and answer a scientific question of their choosing. In this case, the apparatus is a radioactive Cs source, a Geiger-Muller tube with event counting hardware and software, a box of barriers of various material and thickness, and a stand for all of the above with slots in which to place the source and barriers.

The rest of the 200-level labs are NOMR labs documented in Sec. III A: Charge and Mint (VR1 + 2) and the Manifold Lab (VR3–VR6).

C. Data collection

1. Lab epistemology survey and physics identity survey

The LES and PhIS were administered as part of the same survey in all cases.

100-level students completed the presurvey in week 2, after lab 0, which was a traditional hands-on lab, and before VR1, the first NOMR lab. The survey was included as part of a timed quiz, and we recognize the time constraint may have influenced students' responses. The postsurvey was included as part of an untimed reflection on the performance of their group. It was administered in week 4 after the conclusion of VR3, the final NOMR lab for 100-level students. Therefore, the pre-post shifts reported here reflect changes in 100-level students' responses before and after only the NOMR labs.

200-level students completed the presurvey in week 1 before the start of classes and completed the postsurvey in week 10, after their final presentations. The pre-post shifts reported here reflect changes in students' responses before and after all 9 weeks of labs, including non-NOMR activities.

2. Flow survey

The FS was administered every week at the end of the lab to both populations. We offered a small amount of extra credit for each week the survey completed and emphasized that the surveys would help improve the lab curriculum in future terms. In the 100-level population, 42 students completed all eight surveys; in the 200-level population, 14 students completed all eight surveys. The findings reflect responses only from these students who completed all eight surveys.

IV. FINDINGS

A. RQ1: What changes are observed in students' epistemology about experimental physics as a result of the NOMR labs?

Responses to **LES1** and **LES2** were coded independently by the first author and another researcher. The researchers met to reconcile disagreements after the first coding pass, with a disagreement rate of roughly 10% for **LES1** and 30% for **LES2**. After further expanding on the existing code definitions, adding examples, and adjusting

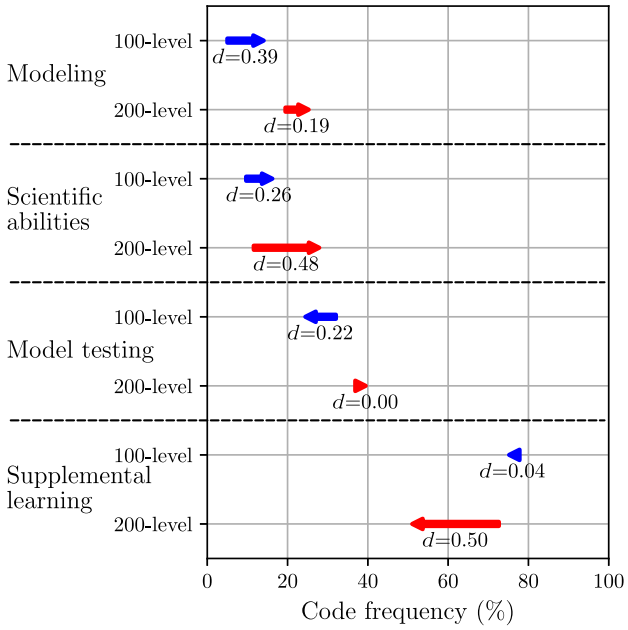


FIG. 3. Code frequencies for both populations' responses to **LES1**. A code frequency represents the percentage of responses in a population that were assigned that code. Each response could be assigned zero, one, or multiple codes, so percentages do not add to 100%. Data from our 100-level population are in blue, and the 200-level population in red; the tail and head of each arrow represent preintervention and postintervention data, respectively. This graphic represents $N_{100} = 278$ and $N_{200} = 38$ matched pre-post responses from 100-level and 200-level students, respectively. Cohen's d is calculated according to Eq. (1).

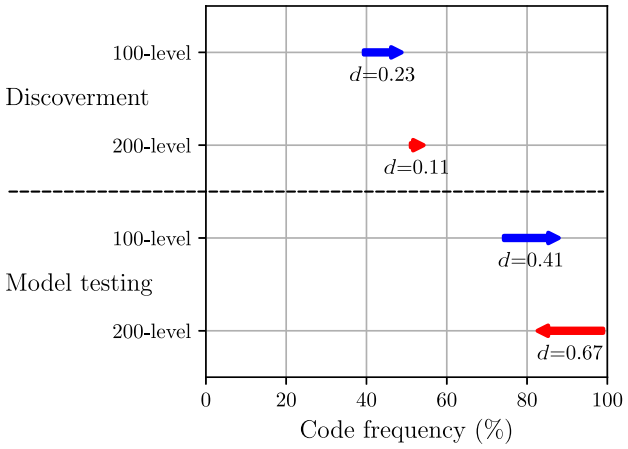


FIG. 4. Code frequencies for both populations' responses to **LES2**. A code frequency represents the percentage of responses in a population that were assigned that code. Each response could be assigned zero, one, or multiple codes, so percentages do not add to 100%. Data from our 100-level population are in blue, and the 200-level population in red; the tail and head of each arrow represent preintervention and postintervention data, respectively. This graphic represents $N_{100} = 278$ and $N_{200} = 38$ matched pre-post responses from 100-level and 200-level students, respectively. Cohen's d is calculated according to Eq. (1).

the codes as described in Sec. II B 1, we reached > 95% interrater agreement across all codes.

The LES findings are presented as code frequencies: the fraction of responses in a population that were assigned each code. Figures 3 and 4 depict the shift from pre to post for each population item. Each student response could be assigned no code, one code, or more than one of the codes associated with the item. As most responses were assigned one or more codes, the total number of codes is greater than the number of responses.

Cohen's d is presented for each shift, calculated from pre- and postcode frequencies p_{pre} and p_{post}:

$$d = \frac{|p_{\text{post}} - p_{\text{pre}}|}{\sqrt{\frac{\sigma_{\text{pre}}^2 + \sigma_{\text{post}}^2}{2}}}; \quad \sigma_i = \sqrt{p_i(1-p_i)}. \tag{1}$$

Whether a code is or is not assigned to a given response is a binary variable, so the binary standard deviation is used.

1. LES1: Why are experiments a common part of physics classes?

Responses to **LES1** were assigned up to four codes, defined with examples in Table II and reproduced below:

Modeling: Experiments in class let students develop their own models for phenomena, discover things on their own, and/or develop their own ideas.

Scientific abilities: Experiments help cultivate students' scientific abilities, such as experimental design, data collection, and data analysis skills.

Model testing: The purpose of doing physics experiments is to prove, support, or test a model.

Supplemental learning: Experiments provide supplemental learning experiences for concepts and theories.

We added the *Modeling* code in response to the recurring presence of its ideas in our dataset and its relevance to our research questions. It captures the creation of new models that result from model-generating experiments in the ISLE process, while *Model testing* represents the subsequent model-testing function of a testing experiment.

Example responses that were assigned to each code are given in Table II. A sample response to **LES1** that was assigned multiple codes follows:

> Experiments allow us to challenge what we know while apply what we have learned. Its a new way of learning things–a more hands-on approach. We can learn about the scientific models and how experiments are designed to either explain a new phenomena or test a pre-existing model.

This response is assigned *Modeling* for the phrase "We can learn… how experiments are designed to either explain a new phenomena…" and *Model Testing* for the last part of

that sentence: "… or test a pre-existing model." *Supplemental learning* is present in a couple of places: "We can learn about the scientific models…" and "Its [sic] a new way of learning things–a more hands-on approach." Finally, the clause "… how experiments are designed…" merits the *Scientific abilities* code.

Both populations' code frequencies are plotted in Fig. 3. Comparing pre-post results, we find that the students in our study became more likely to indicate a belief that labs are meant to develop their *Scientific abilities* and give them opportunities to develop their own models (*Modeling*). The 100-level population became less likely to cite *Model testing* as a purpose of in-class experiments, while the 200-level population became dramatically less likely to cite *Supplemental learning* for the same.

2. LES2: Why do scientists do experiments in professional research?

Responses to **LES2** were tagged with up to two codes, defined with examples in Table II and reproduced below:

Discoverment: Experiments contribute to some aspect of the iterative and generative nature of the scientific process aside from testing an existing model.

Model testing: The purpose of doing physics experiments is to prove, support, or test a model.

LES2's code frequencies are plotted in Fig. 4. Both populations acknowledged the iterative and generative elements of the scientific process (*Discoverment*) more frequently after NOMR labs than at the beginning of the course. The 100-level population's *Discoverment* code frequency increased by a significant degree; the 200-level population's frequency started higher and saw a smaller increase. The 200-level change is small enough to be statistically insignificant.

The 100-level population was assigned *Model testing* codes more frequently after NOMR labs than at the beginning of the course, jumping from 73% to 89%. The 200-level population's code frequencies saw a significant decrease in *Model testing*, moving from 100% to 82% of matched responses.

B. RQ2: What changes are observed in students' physics self-efficacy in experimental physics as a result of the NOMR labs?

1. Statistical analysis

We assess pre-post shifts in students' responses to the Likert items comprising the PhIS by calculating p values using the Wilcoxon signed-rank test [36]. We chose the Wilcoxon test because it is a nonparametric test and thus makes no assumptions about whether the dataset is normally distributed. We opted against using the Mann-Whitney U test as we are interested in testing for differences within paired samples. The Wilcoxon signed-rank test is a more appropriate choice than the Mann-Whitney U test, since the Mann-Whitney test is used for independent samples.

We report the common-language effect size f computed from the Wilcoxon test statistic T and the total rank sum S by the expression: $f = \frac{1+T/S}{2}$. In general terms, f tells us what fraction of students reported a higher score in the postsurvey than in the presurvey. However, this metric does not account for ties, where a student gives the same response for an item in the pre- and postsurveys; ties are ignored in the calculation of the effect size. The common-language effect size can range from 0 to 1 for a given item, where 0 indicates that all respondents gave an equal or lower score to the item on the postsurvey than on the presurvey, 0.5 indicates that as many respondents reported a higher score as reported a lower one, and 1 indicates that every respondent's postsurvey score was equal to or higher than their presurvey score.

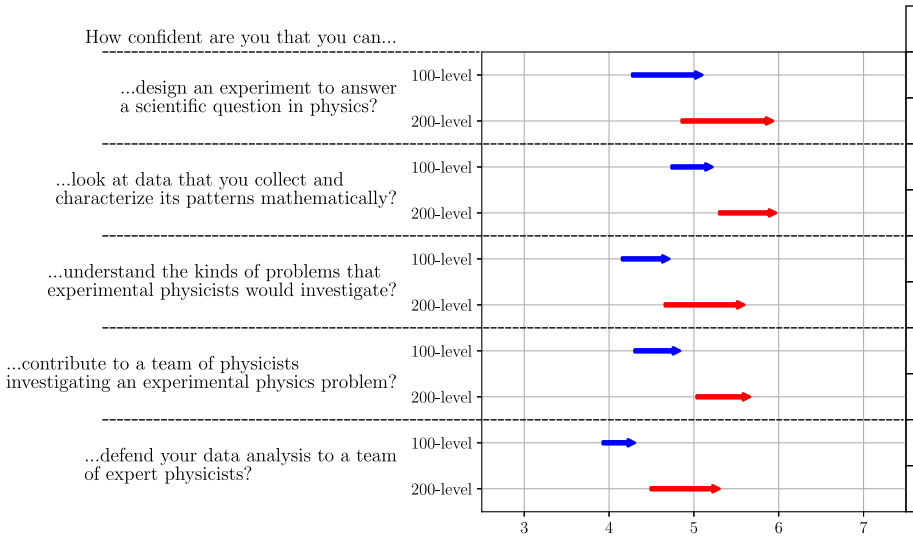
2. Self-efficacy

As shown in Fig. 5, positive shifts are observed for both populations in all self-efficacy items. The figure shows interpolated medians for each item before and after the intervention; each arrow's tail represents the preintervention median, and each head the postintervention median. Thus, the length represents pre-post change. The 100-level and 200-level populations' data are shown in blue and red, respectively. The effect sizes vary, but there is a positive shift for every self-efficacy item at a 99.7% confidence level or better ($p < 0.003$) in both populations.

Comparing the size of the populations' arrows for each item, we note that the shift in the 200-level students' responses is consistently greater than that of the 100-level students, by roughly 50% on average. This magnified effect is observed despite the 200-level predata medians being consistently higher than those of the 100-level students, leaving less room for improvement.

Every individual 200-level response to the item "How confident are you that you can design an experiment to answer a scientific question in physics?" either did not change or became more expertlike. This is a useful example for understanding the common-language effect size. Of all the students whose scores changed, 100% of them increased. Therefore, $f = 1$.

The PhIS results are unique among our dataset in that there are notable differences between responses given by male- and female-identifying students, as shown in Fig. 6. While there are positive shifts for all items for both genders, female-identifying students consistently started lower on each item and saw a smaller increase. Female-identifying students' increases on the last two items are notably small; the shift in responses to **PhIS4** barely meets the traditional significance threshold $p < 0.05$, and the shift for **PhIS5** does not.



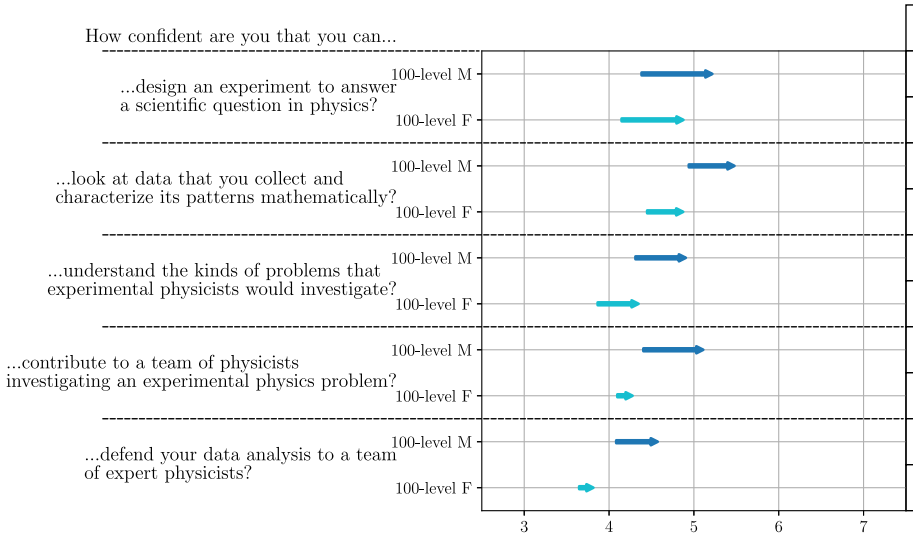
Shift	Effect size f	p-value	T_+
0.890	0.852	2.71e-13	8057.5
1.141	1.000	5.93e-06	351.0
0.545	0.721	1.31e-06	8268.5
0.733	0.925	0.000184	234.0
0.622	0.770	6.02e-09	8154.5
1.000	0.930	0.000157	279.0
0.599	0.699	6.72e-06	9122.0
0.692	0.909	0.000571	230.0
0.443	0.715	2.99e-06	8097.0
0.875	0.842	0.00209	273.5

FIG. 5. PhIS self-efficacy findings for all students who completed both the pre- and postsurveys. 100-level data ($N_{100} = 225$) are in blue, 200-level data ($N_{200} = 35$) in red. Each arrow represents the difference between the interpolated median of the presurvey (tail) and post-survey (head) responses for the associated population item.

C. RQ3: To what extent are students productively engaged in the NOMR activities, and how does that engagement compare with the hands-on labs in the same course?

Following the analysis methods of Karelina *et al.* [28] described in Sec. II B 3, we produced flow plots for each week of each population's lab activities. The number of students who responded with each (x, y) pair, representing (skill/knowledge, challenge) is indicated by the size of the dot at that point. The *area* of each dot is proportional to the number of students it represents. For example, suppose 4 students responded that the lab was extremely challenging ($y = 7$) and they felt only moderately skillful ($x = 4$), and 16 students responded that the lab was a significant challenge ($y = 5$) that they felt prepared to tackle ($x = 5$). We would see a dot at (4, 7) with radius R and a second dot at (5, 5) with radius $2R$. The absence of a dot indicates that zero students gave the associated response.

For our analysis, we invent a quantity to help characterize the quality of student engagement through the lens of flow across an entire class. We calculate the interpolated median of these data along each dimension of the flow plot (skill or knowledge on the x axis and challenge on the y axis) and plot the point defined by these medians. As a



Shift	Effect size f	p-value	T_+
0.899	0.835	6.78e-08	2842.0
0.794	0.874	1.45e-06	1250.0
0.603	0.733	4.31e-05	3414.5
0.490	0.686	0.0157	981.0
0.659	0.737	5.06e-05	3151.0
0.547	0.826	2.44e-05	1138.0
0.771	0.733	5.28e-05	3411.0
0.244	0.638	0.0491	1286.0
0.554	0.763	9.42e-06	3124.5
0.232	0.624	0.0915	1068.5

FIG. 6. PhIS self-efficacy findings for 100-level students who completed both the pre- and postsurveys, divided by gender. Data from male-identifying students ($N_M = 135$) are in dark blue, female-identifying students ($N_F = 88$) in light blue. Each arrow represents the difference between the interpolated median of the presurvey (tail) and postsurvey (head) responses for the associated population item.

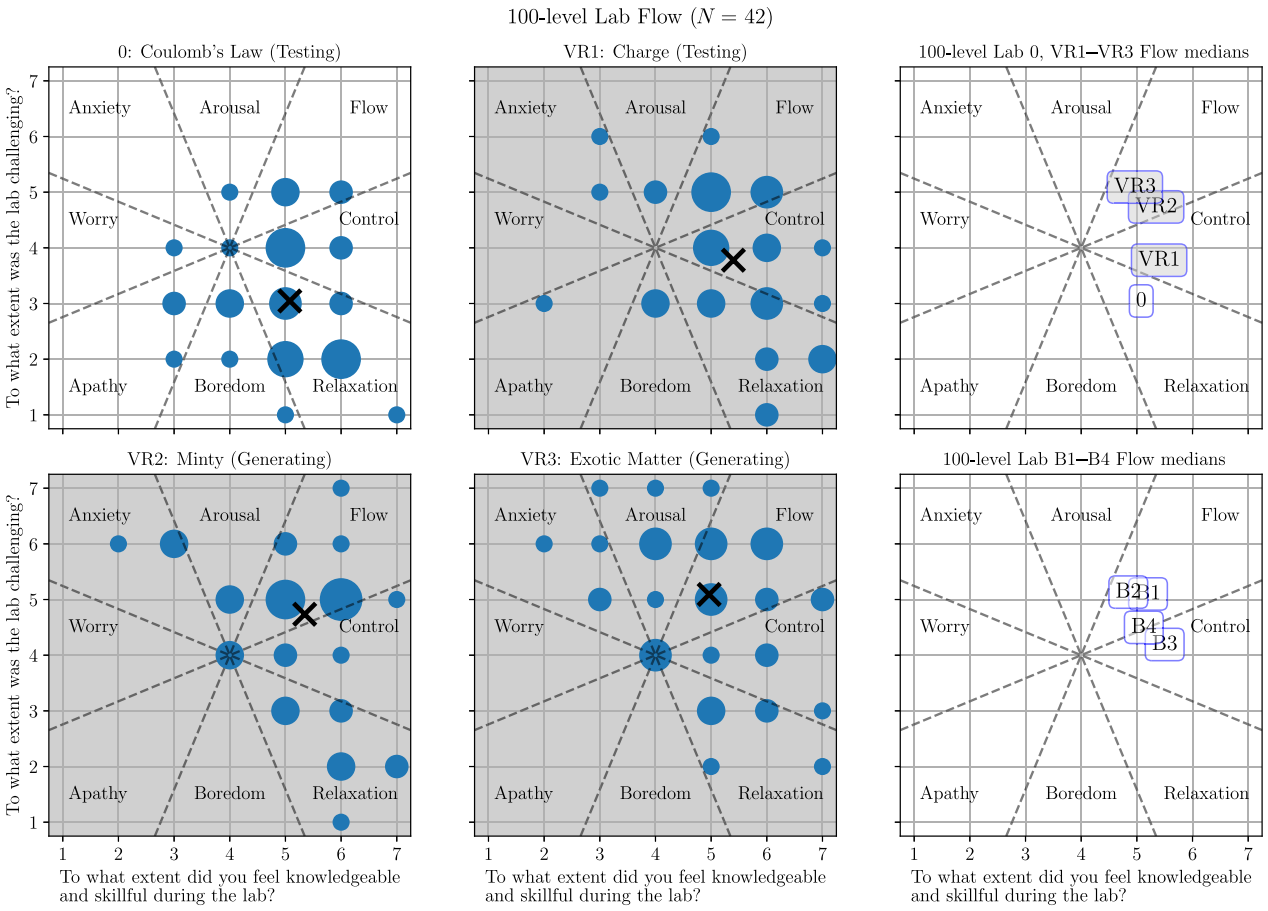


FIG. 7. Flow plots for all $N_{100} = 42$ 100-level students who completed all eight flow surveys. Shaded plots represent VR labs. The black cross on each plot represents the 2D interpolated median of those responses. The area of each dot is proportional to the number of students it represents. The upper right and lower right plots show the medians from the first four and last four labs of the quarter, respectively.

metric for the aggregate engagement state of the class, the closer to the top-right corner this 2D median is, the more effective the lab was at inducing productive engagement. We use the interpolated median rather than the mean, as it better captures the distribution of these ordinal data.

One can produce error bars for the interpolated medians by taking the standard error along each dimension. We found this to consistently produce error bars of the same size or smaller than the marker, so we have omitted them.

The 100-level population's flow data from labs 0–VR3 (the first half of the course) are shown in Fig. 7 alongside summaries of the 2D medians for labs 0–VR3 and B1–B4. The 200-level population's flow data are plotted and their medians summarized in Fig. 8.

We note a few key findings from the flow plots of 100-level students (Fig. 7):

1. In both 100-level model-generating VR labs (VR2 and VR3), zero respondents reported boredom, worry, or apathy. This is not the case in any other 100-level labs.
2. There is gradual migration out of relaxation from labs 0–VR3.
3. The especially high-challenge, high-skill point in the flow channel (6, 6) had zero respondents in the first two labs 0 and VR1, one respondent in lab VR2, and several respondents in lab VR3.

Looking at the migration in the 2D medians of the 100-level students' responses to labs 0–VR3 (Fig. 7, top right), we see a trend consistent with the design of the curriculum: it starts out easy and becomes more difficult by the week (upward movement), but students report feeling equipped to handle the increased difficulty (remain on the right side). Lab VR3, the third and final week of NOMR lab activities, had the median closest to the top-right of the plot out of all of the VR labs.

The rightmost column of Fig. 7 lets us compare the 2D medians of the NOMR labs to the ISLE-based hands-on circuit labs B1–B4, shown on the top-right and bottom-right of the figure, respectively. The students complete all A labs before embarking on the B labs, which comprise the latter half of the course. We note that lab VR3 and the first 2 weeks of circuit labs (B1 and B2) all achieved similar states of productive engagement, landing near the diagonal in the flow channel.

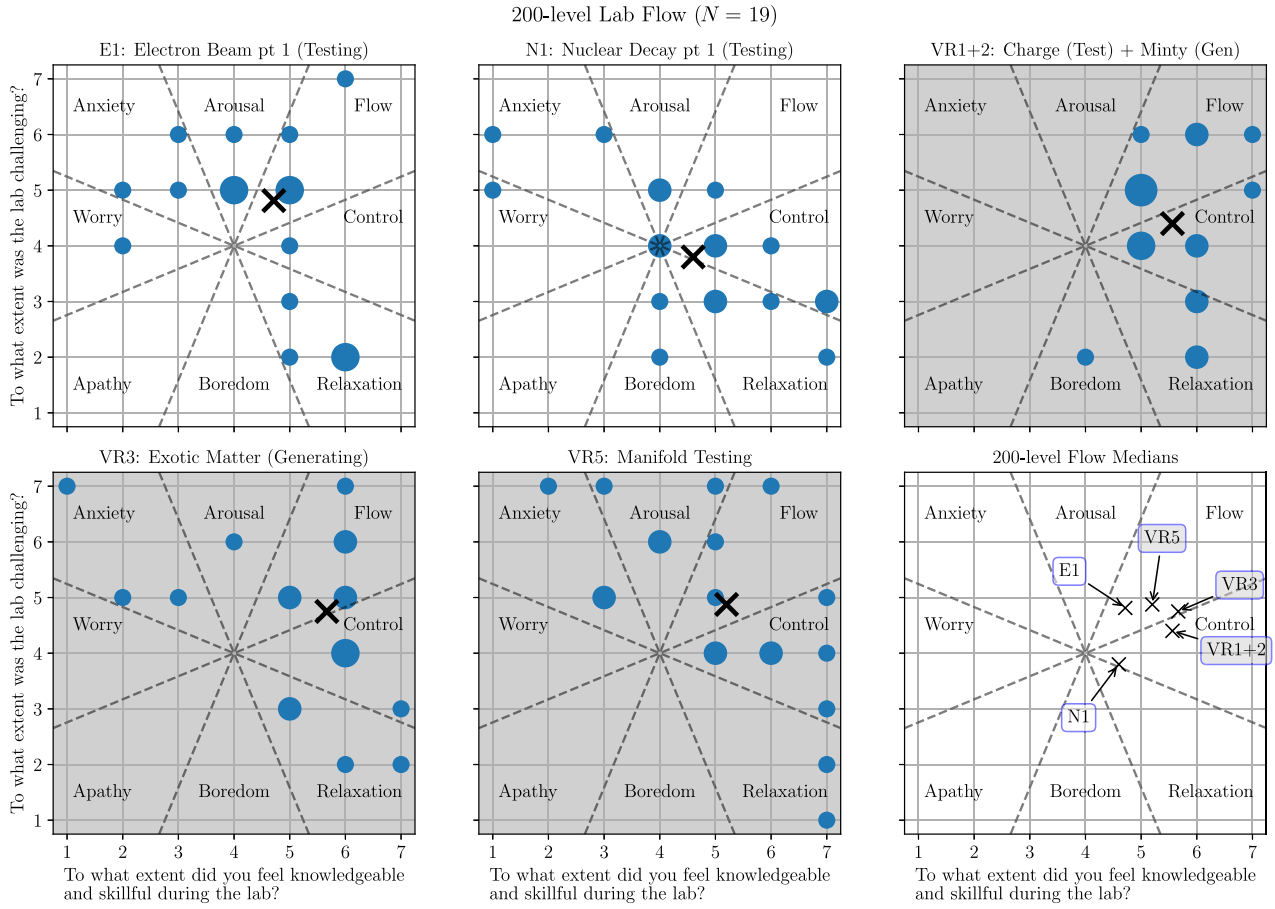


FIG. 8. Flow plots for all $N_{200} = 19$ 200-level students who completed every flow survey. The sixth (lower right) plot shows the medians from the other five plots all together. Shaded plots and labels represent VR labs. The black cross on each flow plot represents the interpolated median of those responses. The area of each dot is proportional to the number of students it represents. Lab activities focusing on communication and writing with no new experimentation, labeled "C" in Table III, are omitted.

In the 200-level population, $N_{200} = 19$ students completed every survey over the course of the quarter; the flow plots and aggregated medians are shown in Fig. 8. Lab activities focusing on communication and writing with no new experimentation (Table III) are omitted from the flow plots. As with the 100-level population, the VR labs' medians follow a path up and to the leftover time, starting in control and landing solidly in flow. The median corresponding to the final VR lab (VR5) is farthest along the diagonal bisecting the flow channel out of all the labs. We note that only three student responses for VR5 are actually in the flow region, with the majority sitting in control and arousal and a few in anxiety and relaxation. The Charge and Mint lab (VR1 + 2) and Exotic Matter lab (VR3) had the most individual responses in the flow region at 7 out of 19.

Figure 9 shows both populations' responses to **F7** ("To what extent was the lab fun and interesting?") for each lab activity. At the 100 level, students' responses are very high for VR1, remain elevated for VR2, and their responses for VR3 are indistinguishable from any other activity.

We see a similar pattern in the 200-level students' responses to the three VR labs in their curriculum: The Charge and Mint lab (VR1 + 2) was seen as the most fun (median ~6.7/7), followed by the subsequent Manifold Lab part A (VR3) with a half-point lower median, and the final Manifold Lab part C (VR5) landed between the hands-on labs with a median of ~5.5/7.

Both populations reported the highest flow state in the last week of NOMR activities. The last NOMR lab for 100-level students was a model generating experiment (VR3). For 200-level students, it was a test of models generated in a prior experiment (VR5).

V. DISCUSSION

A. RQ1: What changes are observed in students' epistemology about experimental physics as a result of the NOMR labs?

1. LES1: Why are experiments a common part of physics classes?

The data show (Fig. 3) that for both populations in our study, the frequency of each code underwent a shift consistent with growth toward an expertlike understanding

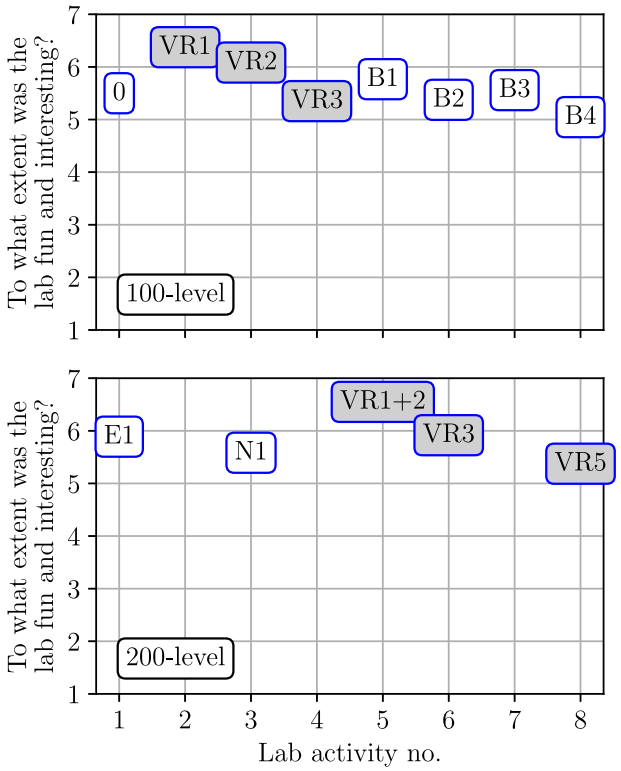


FIG. 9. The top and bottom plots show the 100-level ($N_{100} = 42$) and 200-level ($N_{200} = 19$) populations' interpolated median response each week to **F7** ("To what extent was the lab fun and interesting?"), respectively. NOMR labs are shaded. Lab activities focusing on peer review and writing in the 200-level class are omitted.

of the role of experimentation in physics classes. Both populations started with the *Modeling* code occurring much less frequently than *Model testing*. After the intervention, both received the *Modeling* code more frequently and the 100-level population received *Model testing* less frequently than when they started. In particular, the 200-level population closed the gap entirely: Their postsurvey frequencies for *Modeling* and *Model testing* had equal values at the end of the quarter. The movement toward an equal emphasis on these two codes is considered expertlike as it is in alignment with the epistemological basis of the ISLE approach, itself grounded in the real-world practice of physics.

Supplemental learning came up much more frequently than *Scientific abilities* in pre- and postsurveys in both populations. However, both populations narrowed that gap: Both populations reported *Scientific abilities* more often and the 200-level population reported *Supplemental learning* less often than when they began, which are shifts to more expertlike belief.

The small 100-level *Supplemental learning* shift is a little surprising in the context of the model-generating NOMR labs, in which students are coming up with models for completely fictitious phenomena in a clear and deliberate departure from lecture content. *Supplemental learning* is not a goal of those labs. It could be that the high and unaffected *Supplemental learning* code frequency is due to students' long history of traditional science labs where supplemental learning is indeed the primary goal; years of conditioning are not readily overcome by a 3–4 week intervention.

In sum, The LES code frequencies in both populations reveal that the students' beliefs are changed by the NOMR labs in significant ways. Regarding experimentation as part of their course taking, they are less likely to consider classroom laboratory activities to be a supplement to the theory they learn in lecture (primarily in the form of testing theories they have already learned) and more likely to see them as an opportunity to gain new knowledge in the form of developing scientific abilities and developing scientific models.

2. LES2: Why do scientists do experiments in professional research?

We compare our *Discoverment* code frequencies to the original study's findings by adding together each of their populations' *Model Development* and *Discovery* code frequencies. We expect that a subset of responses in each population received both codes, so these combined *Discoverment* code frequencies are likely overestimates.

We would expect a collection of expert responses to **LES2** to receive both *Discoverment* and *Model testing* codes at a fairly high frequency. The original study's Ph.D. student responses merited *Discoverment* at 65% frequency and *Model testing* at 90% [20]. These were the greatest and smallest frequencies among all populations in the original study, respectively. Ph.D. students were the most experienced population in the original study, suggesting that expertlike change would manifest as an increase in the *Discoverment* code frequency and a decrease in the *Model testing* code frequency.

The increase in *Discoverment* frequencies in both UW populations suggests growth toward an expertlike understanding of the multifaceted role of experimentation in scientific research. On account of the low starting point for both populations, we suggest that we can interpret any increase as developing more toward expertlike beliefs.

The *Model testing* shifts are more ambiguous: the increase in 100-level responses is opposite to what we would consider an expertlike change. However, that increase brings the 100-level *Model testing* frequency almost exactly in line with that of the Ph.D. student population in the original study. On its own, this increase could be generously interpreted to suggest that 100-level students' belief that experiments in scientific research serve to test, support, or prove models was unchanged given the inherent uncertainty in qualitative analysis. More conservatively, this incomplete but accurate belief was bolstered alongside beliefs (i.e., *Discoverment*) that lead to a more

complete expertlike understanding. The 200-level population saw a decrease in *Model testing* codes, moving from 100% to 82% frequency. This is below that of any population in the original study. Unlike the 100-level activities, the 200-level activities—especially the Manifold Lab—are routinely connected to examples of real-world research as part of the introduction for each lab. Explicitly drawing these parallels may have contributed to the relatively large shift in the 200-level responses.

Taken as a whole, these findings represent growth toward an expertlike understanding of the role of experimentation in physics. After NOMR, students shift away from viewing experimental physics exclusively as a theory-testing endeavor, to one that includes a variety of important aspects of the role of experimentation in generating new knowledge. This shift brings students closer to the expert view of scientific knowledge as a process that involves rigorous validation in the natural world.

B. RQ2: What changes are observed in students' physics self-efficacy in experimental physics as a result of the NOMR labs?

The presence of a positive change at a 99.7% confidence level or better ($p < 0.003$) for every PhIS self-efficacy item for both populations suggests that students' self-efficacy around conducting physics experiments is tangibly improved after participating in NOMR labs. Students believe that they are learning in ways consistent with widely agreed-upon undergraduate physics laboratory learning goals [1]. That these shifts are observed in the 200-level population is not especially surprising, as each item represents a core learning outcome for a quarter-long course for physics majors that specifically focuses on experimental physics. It is surprising that we see similar shifts in the 100-level population after just a few weeks of NOMR laboratory exercises. Still, the 100-level shifts' lesser magnitude is consistent with the relatively light depth and duration of the 100-level intervention compared with the 200-level version. All told, students' responses moved closer to those of expert physicists and indicated that their confidence in their own ability to do experimental physics is strengthened significantly.

Looking at the differences by gender in the 100-level data, the lower starting point for female-identifying students across all items is consistent with research into identity and belongingness in introductory physics [23]. Put colloquially, the field of physics is commonly thought of as being an old boys' club, and female-identifying students have on average a harder time developing science identity in physics.

The dramatically smaller shifts in female-identifying students' responses to **PhIS4** and **PhIS5** are of particular interest. These items are focused on one's extrinsic interactions with a group of physicists rather than on one's intrinsic ability to perform a category of tasks. For that reason, it is plausible to believe that these items are inherently gendered; that is, administering these two items would elicit a similar difference by gender in any context. It may be the case that male-identifying students build more confidence in NOMR labs than female-identifying students do, but the absence of gender-distinct results in the LES and flow data suggest the interaction with the headsets seems to be gender neutral. Thus, we hesitate to attribute the results from these items to the instrument or the intervention and highlight this as an area for future study.

Both populations' self-efficacy about designing, conducting, and interpreting experiments is significantly improved after working through NOMR labs. These shifts are aligned with the AAPT laboratory learning objectives [1]. We suggest these data indicate that the NOMR labs are helping students develop confidence in their professional capacity as experimentalists while also helping them develop more expertlike habits of mind about experimental physics.

C. RQ3: To what extent are students productively engaged in the NOMR activities, and how does that engagement compare with the hands-on labs in the same course?

We see the majority of students in the productive zones of flow, arousal, and control during the model-generating NOMR labs. We interpret being in a flow state as optimized student engagement in the learning activities. Achieving flow requires that students know what to do, how to do it, and how well they are doing; students tuning out or becoming lost in the face of the open-endedness of the activities would be reflected in low-skill responses on the left of the flow plots. The data show that the NOMR labs are providing just enough scaffolding to keep students in the zone of proximal development and in flow [27]. Of the first series of 100-level labs (0-VR3), VR3 induced the most productive aggregate state of engagement in students; no responses indicated a state of worry, apathy, or boredom. The 200-level population achieved more productive engagement with the VR labs than either of the hands-on labs E1 and N1.

We recognize that VR is an engaging environment on its own. While it may be hard to disentangle the novelty of VR from the activities themselves, we do see evidence that the novelty wears off. We consider the effect of the gaming/entertainment appeal of VR by examining student responses to item **F7** (plotted in Fig. 9): "To what extent was the lab fun and interesting?" The novelty effect is associated with the introduction of an exciting new technology in the classroom, which induces an initial boost in student engagement that eventually wanes [37].

We estimate that VR's novelty lasts 2 weeks in our context, as we observe in both populations the highest **F7** score in the first NOMR lab, the second highest in the second NOMR lab, and the third NOMR lab is no more or less fun than any of the hands-on labs in the course.

Despite the third week of NOMR labs not benefiting from the novelty effect, students reported the greatest aggregate flow state during that activity (VR3 and VR5 at the 100-level and 200-level, respectively). This suggests that the novelty effect does not fully explain the productive and deep engagement with VR physics labs. Students are not engaged simply because VR is fun; they are engaged because the physics is compelling. NOMR labs use VR specifically because its "secret sauce" of hands-on interaction with fictitious physical phenomena is otherwise impossible. We consider students' strong engagement after the novelty has worn off as evidence that NOMR labs may be leveraging the unique affordances of VR in a pedagogically useful way.

Our comparison of VR and hands-on labs contrasts with Karelina *et al.*'s comparison [28] between students' engagement with video labs and hands-on labs. They found students reported video labs to be slightly more challenging, less fun, and that they felt less skillful when compared to hands-on labs. Our findings demonstrate that students' engagement with VR labs can be similar to or better than their engagement with hands-on labs in the same course. It is important to note that Karelina *et al.* compared two distinct populations of students who went through hands-on and video versions of the same lab activity, while our study compares responses to different lab activities from the same population, so comparisons between our findings and theirs should be made with caution.

We hesitate to overinterpret our analysis of the flow data. In this study, we analyze absolute rather than relative scores. The original application of the eight-channel flow model [32] collected responses from each participant at many points in time over several days. The researchers determined each participant's average response for an item. When creating flow plots, they plotted z scores relative to each participant's average response to each item. This method accounts for the fact that every individual interprets Likert scale questions differently; one person's $3/7$ is another's $6/7$. Karelina *et al.* [28] adapted this original methodology in favor of examining absolute scores due to the limitations of a classroom setting. They had no more than 2–3 responses from any given participant, and we replicated their methodology for comparability.

Further, we note that flow states manifest in neurodiverse learners in ways that are not fully understood [38]. The fact that the measurements of students' flow state take place in a group learning context adds the complexities of group social interactions to the picture. Flow is an individual measurement of an experience that occurs in a group context, which does not give us any information about group dynamics and cohesion. These shortcomings are excellent areas for further work.

D. Use of fictitious physics with virtual reality

Existing survey data do not fully capture the in-class and in-headset experience of NOMR labs. Students' experimental results in NOMR labs are different from reexaminations of well-understood phenomena: Every group's findings are new knowledge students have generated from scratch. We posit that, in this way, NOMR labs may allow students to experience the satisfaction of discovery that professional physicists find so compelling, as reflected in feedback from postcourse surveys:

> I was thrilled and enlightened to be put in a position to analyze physical phenomena that were undocumented and that I had never heard of. Being able to work with a pocket universe and using experimentation to describe it was the best experience in physics courses I've ever experienced. I preferred this VR experience to any physical lab for the sole reason of it being entirely new and having to get every ounce of info about it through experimentation and collaboration.
>
> It's been so much fun learning physics in an exploratory way that focuses on letting us be creative with our thinking. I've not only learned a lot about error analysis and creating models, but also gained a much better perspective on how science and research 'work' in the real world.

As NOMR lab instructors and experienced teachers, we observe students engaging in scientific creativity in ways we have not seen before. Rigorous characterization of creativity is a challenging research task and is not captured in the surveys we administered. We postulate that the use of VR is in part responsible for helping unleash student creativity and highlight this as an area for further study.

This study does not create evidence one way or another about whether VR is necessary for the implementation of the fictitious physics approach to stimulate creativity in introductory labs. We predict that such an intervention would not be as effective without VR, in that the joy of discovery expressed above would likely be altered by the difficulties associated with nonimmersive simulations. Manipulating objects in 3D space on a 2D screen can be challenging, not to mention that it risks widening the conceptual gap between the novel physics and the physics of our universe. Being overly disconnected from a physically interactive 3D space may preclude the suspension of disbelief that allows students to engage so readily in NOMR. It is our experience that the preceding argument is difficult to make convincingly without sharing the in-headset experience, making it rigorous will require substantial human-computer interaction research.

There remains a possibility that a non-VR version's effectiveness could be sufficient for a headset-free version of NOMR to be a favorable tradeoff, considering the expense and overhead associated with VR technology. Presently, Meta Quest 2 headsets (the same used for this study) cost \$300 each; at eight lab groups to a section, the cost of outfitting a classroom to run NOMR labs is ∼\$2400.

While not outrageous, this is beyond the reach of many physics laboratory budgets, especially in underserved communities. Phone-based virtual or augmented reality, 3D simulations experienced on a monitor and controlled via mouse and keyboard or entirely 2D simulations could all employ fictitious physics in a similar way to NOMR labs without the expense of full VR headsets. Further developments based on the work in this paper can contribute to exploring these avenues to open up broader access to this instructional innovation.

E. Epistemological hazards of fictitious physics

The use of fictitious laws of physics raises concerns about whether interacting with fictitious laws of physics can negatively affect students' physical intuition and conceptual understanding. These concerns have been on our minds since the first trial of NOMR in early 2020. For that reason, we take care to maintain conceptual separation between NOMR-unique physics and the physics of our universe:

- The introduction of every NOMR lab manual (except VR1: Charge) makes clear that the physics students will be investigating was created for the purpose of the lab and does not exist in our universe. The lab manuals frame activities with fictitious physics as being original investigations into unknown physics, putting students in the shoes of Coulomb and peers.
- Fictitious particles are given silly names (e.g., minty particles); where they are not named in the lab manual, students are encouraged to come up with their own names for the particles. They discovered them, after all.
- At no point do students work with simulations of both real and fictitious physics at the same time.
- Fictitious particles are only ever referenced in the context of the laboratory component of a course; they are not mentioned in lecture or tutorial components.

To date, we have not seen evidence of negative impacts on students' conceptual or procedural physics knowledge arising from their work with fictitious physics.

VI. CONCLUSIONS

Steering students away from confirmation of known facts and into a different (simulated) universe sounds like a day in the life of Ms. Frizzle's class [39]; in lieu of a magic school bus, we use VR headsets. In both cases, students are transported to the teacher's choice of hands-on learning environment to create knowledge through collaboration with their peers in a fun, engaging, and memorable environment. In NOMR, students learn to gain knowledge in the way that experts do.

This study contributes to the physics education community's effort to create laboratory activities that foster students' growth along affective and epistemological dimensions. We have demonstrated that these goals can be achieved by inventing new physics for students to explore, effectively drawing a new frontier of physics at the introductory level. Our findings (summarized in Table IV) begin to paint a positive picture of the affective impacts of labs featuring fictitious physics, but our experience as instructors suggests that we have yet to capture the full effect of this approach.

Subsequent analysis will include respondent-relative analysis of flow data and examine students' responses across surveys to identify correlations between change toward expertlike beliefs, engagement in lab activities, and development of self-efficacy.

That student engagement remained strong despite the decay of the novelty of VR is an important finding to consider for others interested in the role of immersive technologies in physics education. Replicating traditional

TABLE IV. The research questions, surveys, and outcomes of this study are summarized.

Research question	Survey	Outcome
RQ1: What changes are observed in students' epistemology about experimental physics as a result of the NOMR labs?	LES	Students become more expertlike in their epistemologies associated with the role of experimentation in learning, and in research. NOMR students shift away from viewing experimental physics exclusively as a theory-testing endeavor to one that includes a variety of important aspects of the role of experimentation in generating new knowledge.
RQ2: What changes are observed in students' physics self-efficacy in experimental physics as a result of the NOMR labs?	PhIS	NOMR labs help students develop belief in their professional capacity as experimentalists, while also helping them develop more expertlike habits of mind about experimental physics.
RQ3: To what extent are students productively engaged in the NOMR activities, and how does that engagement compare with the hands-on labs in the same course?	FS	Students become increasingly engaged with successive NOMR labs, even after the novelty wears off. They are most engaged when developing their own hypotheses with novel physics.

or 2D simulator-based physics instructional activities in VR has a track record of being more engaging than comparable treatments in other media but yielding little to no educational benefit in comparison [40–42]. Our findings suggest that instead, educators, researchers, and developers interested in the use of immersive technology in the physics classroom should collaborate to identify niches where its affordances can be leveraged to create unique learning experiences.

Whether VR is strictly necessary to implement the fictitious-physics approach is difficult to say with scientific rigor, and while we do not attempt to do so here, we predict that it is necessary to a greater or lesser degree. That this study found positive results using a VR intervention suggests that the utility of VR in physics education lies in niches where it lets us create learning experiences that would be infeasible or impossible by other means. In other words, an effective educational VR activity is not effective because it uses VR any more than a PhET [43] is effective because it runs in a browser. Rather, it creates a learning experience tailored to solve a specific pedagogical problem, using the medium that provides the most appropriate foundation for doing so.

ACKNOWLEDGMENTS

The authors thank Sabrina Cheng and Tyler Flom for their assistance coding the LES data, Kazumi Tolich for major contributions to NOMR's instructional materials and allowing us to administer all of these surveys in her course, David Aplin and Alexis Mendoza for heroically taking on all of the technical legwork to facilitate NOMR's classroom implementation, and Eugenia Etkina, Peter Shaffer, and Charlotte Zimmerman for their insightful feedback on early drafts of this article. This work would not have been possible without the many TAs who have taught NOMR labs and the supportive community of the UW Physics Education Research Group; both groups are too numerous to list here, but you know who you are. Thank you. We gratefully acknowledge financial support from the UW Student Technology Fee. This material is based upon work supported by the National Science Foundation under Grant No. DGE-1256082.

[1] J. Kozminski, H. Lewandowski, N. Beverly, S. Lindaas, D. Deardorff, A. Reagan, R. Dietz, R. Tagg, M. Eblen-Zayas, J. Williams, R. Hobbs, and B. Zwickl, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum*, Technical Report (American Association of Physics Teachers, College Park, MD, 2014).

[2] E. M. Smith and N. G. Holmes, Best practice for instructional labs, Nat. Phys. **17**, 662 (2021).

[3] N. G. Holmes, J. Olsen, J. L. Thomas, and C. E. Wieman, Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content, Phys. Rev. Phys. Educ. Res. **13**, 010129 (2017).

[4] D. Hu, B. M. Zwickl, B. R. Wilcox, and H. J. Lewandowski, Qualitative investigation of students' views about experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 020134 (2017).

[5] E. Etkina and A. V. Heuvelen, Investigative science learning environment—A science process approach to learning physics, in *Research-Based Reform of University Physics* (American Association of Physics Teachers, College Park, MD, 2007), Vol. 1.

[6] A. Bandura, Self-efficacy: Toward a unifying theory of behavioral change, Psychol. Rev. **84**, 191 (1977).

[7] E. Etkina, A. Van Heuvelen, S. White-Brahmia, D. T. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 020103 (2006).

[8] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

[9] R. Trumper, The physics laboratory—A historical overview and future perspectives, Sci. Educ. **12**, 645 (2003).

[10] D. T. Brookes, E. Etkina, and G. Planinsic, Implementing an epistemologically authentic approach to student-centered inquiry learning, Phys. Rev. Phys. Educ. Res. **16**, 020148 (2020).

[11] E. Etkina, D. T. Brookes, and G. Planinsic, The investigative science learning environment (ISLE) approach to learning physics, J. Phys. Conf. Ser. **1882**, 012001 (2021).

[12] E. Etkina, Millikan award lecture: Students of physics—Listeners, observers, or collaborative participants in physics scientific practices?, Am. J. Phys. **83**, 669 (2015).

[13] E. Etkina and G. Planinsic, Defining and developing 'critical thinking' through devising and testing multiple explanations of the same phenomenon, Phys. Teach. **53**, 432 (2015).

[14] E. Etkina, A. Warren, and M. Gentile, The role of models in physics instruction, Phys. Teach. **44**, 34 (2006).

[15] M. M. Stein, E. M. Smith, and N. G. Holmes, Confirming what we know: Understanding questionable research practices in intro physics labs, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Stein.

[16] J. P. Canright, J. R. Olsen, and S. W. Brahmia, Leveraging virtual reality for student development of force models in the introductory lab, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.Canright.

[17] J. P. Canright and S. White Brahmia, Developing expertlike epistemologies about physics empirical discovery using

virtual reality, presented at PER Conf. 2021, virtual conference, 10.1119/perc.2021.pr.Canright.

[18] 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, Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[19] M. Vignal, G. Geschwind, B. Pollard, R. Henderson, M. D. Caballero, and H. J. Lewandowski, Survey of physics reasoning on uncertainty concepts in experiments: An assessment of measurement uncertainty for introductory physics labs, Phys. Rev. Phys. Educ. Res. **19**, 020139 (2023).

[20] D. Hu and B. M. Zwickl, Examining students' personal epistemology: The role of physics experiments and relation with theory, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.juried.003.

[21] 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).

[22] J. M. Reilly, E. McGivney, C. Dede, and T. Grotzer, Assessing science identity exploration in immersive virtual environments: A mixed methods approach, J. Exp. Educ. **89**, 468 (2021).

[23] Z. Hazari, P. M. Sadler, and G. Sonnert, The science identity of college students: Exploring the intersection of gender, race, and ethnicity, J. Coll. Sci. Teach. **42**, 82 (2013), https://www.jstor.org/stable/43631586.

[24] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).

[25] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience* (Harper & Row, New York, NY, 1990).

[26] M. Csikszentmihalyi, *Applications of Flow in Human Development and Education: The Collected Works of Mihaly Csikszentmihalyi.* (Springer Science + Business Media, New York, NY, 2014), pp. 494, xxii, 494–xxii.

[27] N. S. Rebello and D. Zollman, Problem solving and motivation—Getting our students in flow, presented at PER Conf. 2013, Portland, OR, 10.1119/perc.2013.inv.008.

[28] A. Karelina, E. Etkina, P. Bohacek, M. Vonk, M. Kagan, A. R. Warren, and D. T. Brookes, Comparing students' flow states during apparatus-based versus video-based lab activities, Eur. J. Phys. **43**, 045701 (2022).

[29] V. K. Zaretskii, The zone of proximal development, J. Russ. East Eur. Psychol. **47**, 70 (2009).

[30] D. L. Schwartz, J. D. Bransford, D. Sears *et al.*, Efficiency and innovation in transfer, in *Transfer of Learning from a Modern Multidisciplinary Perspective* (Information Age Publishing, Charlotte, NC, 2005), Vol. 3, p. 1.

[31] F. Massimini, M. Csikszentmihalyi, and M. Carli, The monitoring of optimal experience: A tool for psychiatric rehabilitation, J. Nerv. Men. Dis. **175**, 545 (1987).

[32] F. Massimini and M. Carli, *Optimal Experience: Psychological Studies of Flow in Consciousness*, edited by M. Csikszentmihalyi and I. S. Csikszentmihalyi (Cambridge University Press, New York, NY, 1988), pp. 266–287.

[33] S. A. Jackson and H. W. Marsh, Development and validation of a scale to measure optimal experience: The flow state scale, J. Sport Exercise Psychol. **18**, 17 (1996).

[34] J. Day, J. B. Stang, N. G. Holmes, D. Kumar, and D. A. Bonn, Gender gaps and gendered action in a first-year physics laboratory, Phys. Rev. Phys. Educ. Res. **12**, 020104 (2016).

[35] N. Holmes, G. Heath, K. Hubenig, S. Jeon, Z. Y. Kalender, E. Stump, and E. C. Sayre, Evaluating the role of student preference in physics lab group equity, Phys. Rev. Phys. Educ. Res. **18**, 010106 (2022).

[36] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. **1**, 80 (1945).

[37] R. E. Clark, Reconsidering research on learning from media, Rev. Educ. Res. **53**, 445 (1983).

[38] A. McDonnell and D. Milton, Going with the flow: Reconsidering 'repetitive behaviour' through the concept of 'flow states', in *Good Autism Practice: Autism, Happiness and Wellbeing*, edited by G. Jones and E. Hurley (BILD, Birmingham, UK, 2014), pp. 38–47.

[39] J. Cole and B. Degen, *The Magic School Bus Lost in the Solar System* (Scholastic, New York, NY, 1990).

[40] J. R. Smith, A. Byrum, T. M. McCormick, N. Young, C. Orban, and C. D. Porter, A controlled study of stereoscopic virtual reality in freshman electrostatics, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.089.

[41] C. D. Porter, J. R. Brown, J. R. Smith, E. M. Stagar, A. Simmons, M. Nieberding, A. Ayers, and C. Orban, A controlled study of virtual reality in first-year magnetostatics, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Porter.

[42] J. H. Madden, A. S. Won, J. P. Schuldt, B. Kim, S. Pandita, Y. Sun, T. J. Stone, and N. G. Holmes, Virtual reality as a teaching tool for moon phases and beyond, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Madden.

[43] K. Perkins, W. Adams, M. Dubson, N. Finkelstein, S. Reid, C. Wieman, and R. LeMaster, PhET: Interactive simulations for teaching and learning physics, Phys. Teach. **44**, 18 (2006).