Developing model-making and model-breaking skills using direct measurement video-based activities

Matthew Vonk,[1,*] Peter Bohacek,[2] Cheryl Militello,[†] and Ellen Iverson[3]

[1]*University of Wisconsin River Falls, River Falls, Wisconsin 54022, USA*
[2]*Henry Sibley High School, 1897 Delaware Ave Delaware Avenue,*
Mendota Heights, Minnesota 55118, USA
[3]*Science Education Resource Center at Carleton College (SERC),*
One North College St., Northfield, Minnesota 55057, USA

(Received 11 July 2016; published 11 August 2017)

This study focuses on student development of two important laboratory skills in the context of introductory college-level physics. The first skill, which we call model making, is the ability to analyze a phenomenon in a way that produces a quantitative multimodal model. The second skill, which we call model breaking, is the ability to critically evaluate if the behavior of a system is consistent with a given model. This study involved 116 introductory physics students in four different sections, each taught by a different instructor. All of the students within a given class section participated in the same instruction (including labs) with the exception of five activities performed throughout the semester. For those five activities, each class section was split into two groups; one group was scaffolded to focus on model-making skills and the other was scaffolded to focus on model-breaking skills. Both conditions involved direct measurement videos. In some cases, students could vary important experimental parameters within the video like mass, frequency, and tension. Data collected at the end of the semester indicate that students in the model-making treatment group significantly outperformed the other group on the model-making skill despite the fact that both groups shared a common physical lab experience. Likewise, the model-breaking treatment group significantly outperformed the other group on the model-breaking skill. This is important because it shows that direct measurement video-based instruction can help students acquire science-process skills, which are critical for scientists, and which are a key part of current science education approaches such as the Next Generation Science Standards and the Advanced Placement Physics 1 course.

DOI: 10.1103/PhysRevPhysEducRes.13.020106

I. INTRODUCTION

Being able to explain a complex natural phenomenon with a simple model is powerful, elegant, and useful [1,2]. It is an ability that goes straight to the heart of what it means to be a scientist. Yet most high school students never get a chance to put it into practice since they are exposed to only a limited range of laboratory activities and even those do not often help them to "fully understand science process." [3] Many leading textbooks circumvent this problem by simply providing students with the equations that they need. While giving students the neatly packaged results of other people's work may seem like an efficient approach, we feel it leaves out some important steps and seems (from our constructivist mindset) to be like giving students the punch line without ever telling them the joke.

A. A summary of the literature on model making

In the late 1970s Hestenes strongly advocated that teachers help students develop their own models. [4]

> *…mathematical modeling should be the central theme of physics instruction. This means that the teaching of physical facts and theories should be subsidiary to teaching the principles and techniques of mathematical modeling* [5].

Research has shown that students who are instructed using the modeling method outperform students in traditional classes [6]. In the subsequent decades, Hestenes' work grew into the Modeling Instruction program, an international movement that remains vital and productive today [7].

While Modeling Instruction has reached 10% of teachers at the secondary level [8], other researchers have popularized similar ideas at the undergraduate level. For decades, McDermott has been working to promote the practice of allowing students to learn physics through inquiry.

> *Students are guided through carefully sequenced activities and questions to make observations that they can use as the basis for their model.* [9]

[*] matthew.vonk@uwrf.edu
[†] Present address: 29 Kilburn Rd Garden City, New York 11530, USA.

More recently, Brewe expressed the need for modeling by pointing out that models allow students to "see science as a process and scientific knowledge as a work in progress." [10]

Similarly Windschitl, Thompson, and Braaten champion the use of inquiry learning, but warn that too often student-directed investigations are implemented apart from a grounding conceptual model (e.g., when students test how the growth of a plant is affected by music, Coca-Cola, or being upside down). "Because such questions are arbitrary—i.e., make no sense without the context of at least a beginning model for understanding the phenomenon—then any hypotheses emerging from these questions are likely to be little more than poorly informed guesses." [11]

Although previous studies have measured the efficacy of the modeling process as a whole [2,6,9], and other studies have measured how modeling affects performance on specific tasks (such as experimental design [12], trouble shooting [1], and explaining phenomena [13]) for this study we opted to focus on two smaller, but important components of the modeling process.

B. A review of the literature on model breaking

Physicists are famous for their simplified models, like the frictionless surfaces and massless pulleys that tend to show up in introductory classes. And, while it may make sense to start there, most introductory classes rarely move beyond the idealized cases. Yet, we do our students a disservice when we gloss over perturbations that are plainly obvious to students and compromise the validity of the model. Our sanitized reality too often leaves students with the impression that physics does not actually apply to the real world [14].

Brewe contrasts traditional instruction where the "content is permanent, [and] all validation has already taken place" with modeling instruction, where "models are temporal, [and] must be validated, refined, and applied [10]." An important step toward critically evaluating models is being able to determine if a model matches the data. This fundamental step can be quite challenging for introductory students. In order to make that judgment, students must be able to characterize the uncertainty in their measurements and calculations.

In 1969 Bevington and Robinson published their seminal work, *Data Reduction and Error Analysis for the Physical Sciences* [15]. While this was a valuable resource, students continued to struggle with these skills, so that nearly three decades later Allie and Buffler and their collaborators [16] could publish a study of first-year university students that showed that their "intuitive ideas about improving precision and accuracy of measurement are not distinct" and that their "procedural understanding is context dependent".

In 2001, Deardorff found [17] that students often, "make arbitrary judgments about the agreement between results and fail to consider the uncertainty estimates when making these comparisons." While this is an unfortunate result, it is not a surprising one.

On a more positive note, several researchers have reported making significant strides in changing the way that their students think about measurements and data. In 2005, Kung reported [18] that the fraction of her students that used "range and not just average when comparing two data sets approximately doubled after instruction." More recently, in 2015 Holmes, Wieman, and Bonn conducted a semester-long study that scaffolded students through the process of quantifying their experimental uncertainty and using that uncertainty to target optimal areas for experimental improvements. They reported [19] that, "Students in the experimental condition were 12 times more likely to spontaneously propose or make changes to improve their experimental methods than the control group." In addition, "the students in the experimental condition were also four times more likely to identify and explain the limitation of a physical model using data." "The differences between the groups were seen to persist into a subsequent course taken the following year."

C. A review of the literature on student-analyzed video

The area of student-analyzed videos and other web-based resources has also been an active area of research. A meta-study indicates that students using instruction that includes a significant online component do "modestly better, on average than those learning through traditional face-to-face instruction [20]." However, elements such as video do not influence the amount that students learned unless the learners were given "control of their interactions with the media" or prompted to reflect on their activities. [20] Since the early 2000's, the LivePhoto group (Teese, Laws, Sokoloff, and many others) has pioneered exactly this type of *interactive* web-based video analysis [21]. The project has provided free online video recordings of real events that students can analyze for themselves. The learning activities associated with these videos offer advantages compared to other methods of instruction:

- Compared to labs, they do not require expensive and time-consuming physical equipment.
- Compared to passive lectures they allow the students to actively collect data and analyze it for themselves.
- Compared to simulations, they depict real events and not idealized computer models of events [22].
- The web-based distribution of the videos makes them suitable for online courses, distance learning, flipped classes and labs, and homework.
- Their on-demand availability means that they can be used by students who missed class or who simply require more time to learn.

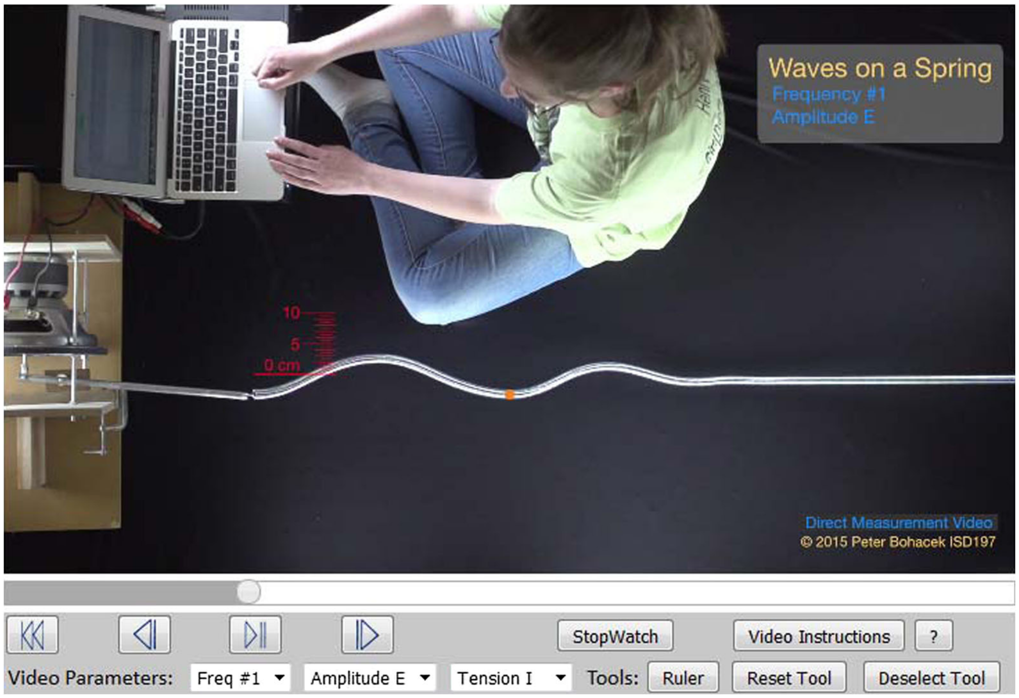
A decade after starting the LivePhoto project, LivePhoto founders Teese and Laws joined with Koenig to reinvent the project as Interactive Video Vignettes [23,24] or IVVs. The IVVs include integrated instructional content with the video and provide a mechanism for student response. That said, most of the IVVs present a single linear narrative arc that does not give students the freedom to explore scenarios of their own choosing. Despite that constraint, IVVs have been shown to improve student performance on the Force Concept Inventory [25]. While their results certainly inform our work, for this study we were interested in exploring the impact of video-based educational resources on laboratory skills.

Another project that has used student-analyzed video and student-directed investigation is ISLE, which stands for Investigative Science Learning Environment. ISLE was developed by Etkina, and Van Heuvelen, and it has since grown to include many other collaborators. The ISLE approach leads students through the process of making careful observations, proposing tentative explanations, and then devising experiments to rule out incorrect explanations. Once an explanation has proved reliable, it is applied to a new situation. The process is iterative and cyclical, “at any step, one can go back and revisit the previous step or examine the assumptions [26].” “The goal is to engage students in actions and decisions similar to those of real physicists by working with simple experiments [27].” The ISLE method has been shown to help students design experiments, analyze data, and communicate like scientists [12,28,29]. Although the ISLE website [30] hosts many ISLE video experiments, the ISLE research thus far has concentrated on the efficacy of apparatus-based (rather than video-based) learning.

Although there has been much work dedicated to fostering students’ modeling skills and further work geared toward exploiting the advantages of student-analyzed video, the work has remained largely separate (the ISLE group is a notable exception). This is likely because the technology simply did not exist until recently to allow students to vary multiple parameters in a single video scenario. The advent of direct measurement video (DMV) matrices has opened up a completely new field to explore.

II. DIRECT MEASUREMENT VIDEOS

The traditional opportunity for students to hone their science skills is in the laboratory. Unfortunately, many lab activities are formulaic and do not actually prepare students for physics exams [31] let alone for the open-ended investigative endeavors that we typically associate with the practice of science [32]. Perhaps an even larger limitation of the traditional laboratory is that student access to physical apparatus is often restricted to brief lab periods. That means that student learning of critical science skills is also restricted. If scientific abilities can only be learned in a laboratory setting, then other potential opportunities (such as learning outside of the classroom, during “lecture” classes, and via online classes) may go unrealized.



Video Parameters: Freq #1 Amplitude E Tension I

FIG. 1. This video shows a wave moving down a large horizontal spring. Students can access interactive tools by selecting the *Ruler* and *Stop Watch* buttons on the bottom toolbar. Using the tools, they can measure the amplitude, frequency, wavelength, period, and speed of the wave. As indicated on the bottom toolbar the current video shows the lowest of five possible frequencies, the largest of five possible amplitudes, and the lowest of three possible tensions. By using the drop-down menus students can select different values for these parameters, which allows them to easily perform a range of experiments. For the model-making assessment (Appendix A), students in classes that had not studied wave phenomena were asked to determine the relationship between the wavelength and frequency of a wave. Successfully completing that task required students to design, execute, and analyze an experiment where the frequency of the wave was varied and the wavelength was measured (while holding the other parameters constant). You may use this online resource for free at the link in Ref. [35].

Another way for students to observe phenomena, make measurements, collect, and analyze data is using direct measurement videos [33]. DMVs are short high-quality videos that show a scientifically interesting event. Students are able to analyze the event using provided online tools (rulers, protractors, stopwatches, etc.). Some of the DMVs are published as single videos [34], but others are published as an interconnected web or matrix of videos that the user can navigate using an integrated console (see Fig. 1). This gives the user a sense that they are actually controlling what happens in the video. For example, the wave properties video matrix [35] shows a 1.8-meter horizontal spring being oscillated by a hefty subwoofer stereo speaker. From a single video, students can measure the frequency, wavelength, amplitude, period, and velocity of the waves. But then, students can *change* what is depicted in the video to see how that affects the other parameters. In the case of the wave properties matrix, the user can select between five different frequencies, five different amplitudes, and three different tensions. These options create a $5 \times 5 \times 3$ matrix containing 75 videos in all. That freedom allows students to design and execute experiments to answer questions that they generate themselves in much the same way that they would in an inquiry lab with physical apparatus. For example, one group of students might want to determine if the speed of the wave increases with frequency. Another group of students in the same class might be curious to see if the wave slows down as its amplitude decreases. A third group might want to determine the quantitative relationship between the wavelength and the frequency of a wave. All three groups could then design an experiment using the same Wave Properties video matrix that would allow them to collect data, analyze it, and ultimately answer their question.

DMVs have some decided advantages over physical apparatus. For one thing, student access to DMVs is not restricted in the same way that it is for physical apparatus. Instead, DMVs are available on demand to any student with an internet enabled device. In addition, DMVs offer an interesting blend of freedom and constraint. The user is free to make many decisions about what and how to measure and (with the video matrices) even *what* happens in the video, but the video is also constrained in ways that may help to keep students on track and help them to focus on the essential physics of the interaction.

While existing DMVs are not well suited for *all* lab practice skills (e.g., learning to use or to debug lab equipment, learning laboratory safely, or learning to perform experimental design using physical apparatus) we were curious to determine if they could be an effective lab supplement for other important science skills. The two skills that we chose to study were model making and model breaking because they are critical parts of the scientific process and they can be difficult to convey outside of a laboratory setting.

III. DESIGNING THE STUDY

A. What do we mean by model making?

The modeling literature generally describes the modeling process as a cyclic iteration of a number of component steps. While virtually all researchers include the process of coming up with a model, there is no consensus on what to call this step. Thus many terms are used, including *constructing* [1,13,36,37], *developing* [6,11,37], *generating* [11], *building* [11], and *creating* [11] a model. Others describe the process as "drawing inferences… [and] formulating… hypotheses" [9]. We have developed a list of component abilities that captures the essence of this creative step in the modeling process and maps easily to a practical rubric. We have chosen to call this particular set of skills *model making*.

We define model making as the ability to

- devise an experiment to determine a quantitative relationship between different variables,
- perform the experiment and collect data [38],
- analyze the data in a way that leads to the construction of a quantitative model,
- apply the model by using it to express the relationship graphically, algebraically, and verbally, and
- use the model to make predictions.

In order to create a model-making rubric, we looked to two primary sources. The first, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* [39], details several relevant focus areas including modeling, designing experiments, analyzing and visualizing the data, and communicating physics.

The second document that played a foundational role in our attempts to quantify the model-making skill was the 2014 paper published by Zwickl, Finkelstein, and Lewandowski [36]. It provides a very useful graphical schematic of the modeling process (which we have included in a slightly modified form in Fig. 2). The article details the steps involved in "constructing models, using them to make predictions and explanations, using them to interpret data, comparing predictions and data, and refining models based on new evidence."

There are two major differences between the way that we framed the model-making skill and the framework presented in the Zwickl paper, and both differences have to do with the fact that their paper was focused on an advanced lab course whereas our study was conducted at the introductory level. The first difference is that their diagram emphasizes the measurement model (the left half of Fig. 2) in addition to the physical system model. While it is important for advanced students to understand the measurement tools they are using, this aspect is "often minimized as a learning goal at the introductory level [36]".

The second difference is that the Zwickl paper indicates that "the primary goal of modeling in upper-division lecture and lab courses is usually not to uncover new basic

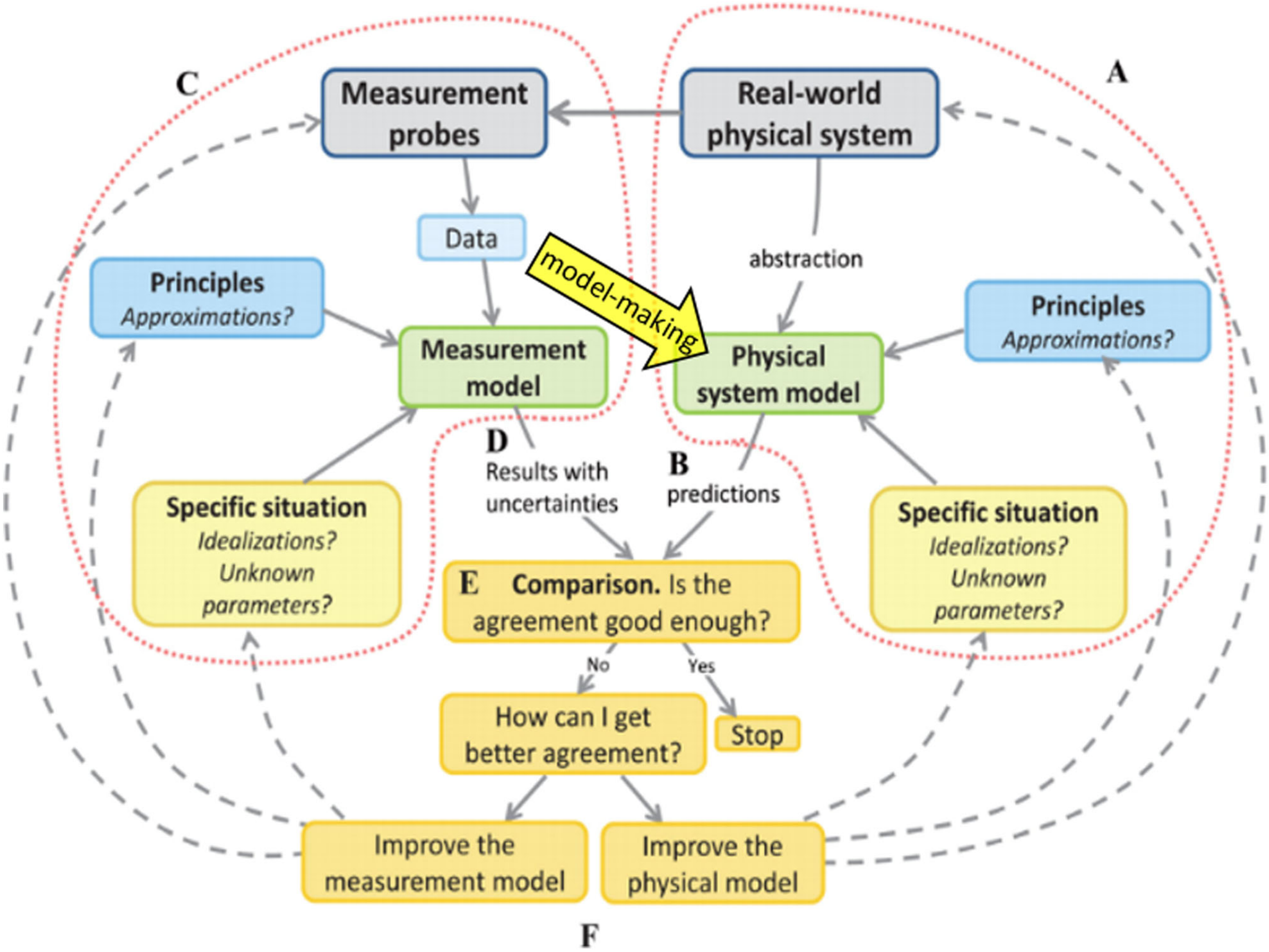


FIG. 2. This graphic is a slightly modified version of a figure from Zwickl, Finkelstein, and Lewandowski's paper on modeling [36]. The diagram does an excellent job of concisely representing the steps in the process of modeling in science. Within this framework, model breaking is represented as a sequence of steps that begins with *Data* (upper left) and ends at stage E. *Note: for this work we did not emphasize the intermediate step "measurement model" which appears in that sequence of steps.* Our conception of model-making (going from data to a physical system model) did not appear as a transition in the original graph, so the diagram was modified by the addition of the yellow "model-making" arrow. It should be noted that a modified version of the diagram has recently been published [1], and further modifications are in process [private communication]. Here is the figure's original caption: *A framework for modeling in a physics laboratory involving (A) construction of a model of a real-world physical process, (B) making predictions about the behavior of the physical system, (C) creating a model of measurement tools, (D) using the measurement model to interpret the data and understand the limitations of measurements, (E) make comparisons between the data and predictions, and (F) model refinement.*

principles… [but rather] models are constructed by applying known principles (e.g., Maxwell's equation) to describe a specific, possibly complex, real-world phenomenon." Since our work is done at the introductory level, we absolutely want our students to uncover new basic principles using data that they collect themselves. For that reason we modified the diagram (see Fig. 2) to include that step. Equipping students to find that path for themselves is a key aspect of the model-making skill as we defined it.

B. What do we mean by model breaking?

Virtually all educational researchers who discuss the modeling process include a step where the model is critically compared to observations in order to determine if the model accurately describes the data. Many terms are used for this process including *evaluating* [11], *testing* [9,11,26], *refining* [10], *comparing* [1,36], and *revising* [40].

While each of these words gets close to describing the skill we were interested in measuring, none contained the exact mix of component skills that we were looking for. This is especially true of the words *refining* and *revising* which include the process of adjusting the model if it is found to deviate from reality. While the ability to make those adjustments is important, it was beyond the scope of this study.

In order to create a model-breaking rubric we looked to the work of Holmes [40]. This led us to define model breaking as the ability to critically evaluate if the behavior of a system is consistent with a given model. This ability was broken down into a number of component skills:

- estimating the absolute uncertainty in a measurement,
- calculating the relative uncertainty of a number,
- propagating absolute (relative) uncertainties when adding (multiplying) two numbers,
- adding uncertainties as the square root of the sum of squares, and
- referencing calculated measurement uncertainties when evaluating the correspondence of the data to a model.

IV. METHODS

A. Participants and activities

The goal of this study was to measure student mastery of model-making and model-breaking skills. It involved one class section of *Algebra-Based Introductory Physics* and three sections of *Calculus-Based Introductory Physics*. All of the classes were taught in a SCALE-UP [41] format that met four times a week, twice for 110 min and twice for 50 min. Each section had a different instructor. The total enrollment of all four sections was 116 students.

Students were divided into treatment groups (model making and model breaking) using matched-pair random assignments based on Force-Motion Concept Evaluation (FMCE) pretest scores. Throughout the semester, students completed five DMV-based activities [42], with each treatment group working on an activity that was specially scaffolded for their skill. All of the students worked with one or two partners on the activities and on the summative assessment. The first two activities were performed during a short period (50 min), but many students were not able to complete the activities so we switched to the long period (110 min) for the last three activities and the final assessment.

The scores for two neutral assessment items were also used to establish equivalency between the treatment groups. These control questions asked students to calculate the initial momentum of the disk, and the final momentum of the disk and cart (see Fig. 3) but had no model-making or model-breaking aspects. Groups that correctly measured the initial momentum to within 10% of the correct value received one point, while groups that measured the initial momentum between 10% and 25% received a half point. Groups whose deviation from the correct value exceeded 25% received no credit. Points for determining the final momentum were awarded in the same way.

An additional instructor of two algebra-based sections was asked to participate in the study but during the course of the semester, several students in those sections switched treatment groups. The completed assessments from those sections were used as practice packets to train the three scorers to consistently apply the rubric, but those results were then excluded from all analyses.

B. Assessment

Near the end of the semester, we administered a single 2-h assessment of both skills to both treatment groups. Existing lab groups (usually 2 or 3 people) collaborated on the assessment, turning in one assessment per group. There were 23 lab groups (58 individuals) in the model-making treatment group and 24 lab groups (58 individuals) in the model-breaking treatment group.

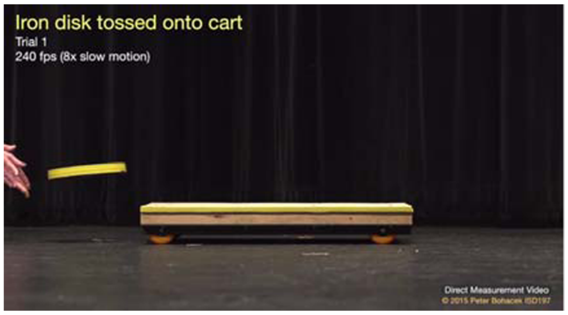


FIG. 3. An iron disk is tossed onto a cart where it bounces and slides. For the model-breaking assessment (Appendix B), students were asked if the total momentum of the disk and cart before the collision was the same as their combined momentum after the collision despite the obvious messiness of the collision itself [43].

The model-making part of the assessment (Appendix A) was based on the wave properties video matrix (Fig. 1) [35], which allows users to independently vary the amplitude, frequency, and tension in a large horizontal spring. Students were asked to, "Use measurements from the videos to determine the mathematical relationship between the wavelength of a wave and the frequency of that wave." By the time the assessment was administered the model-making group had worked on five model-making activities, but *none* of the classes had ever studied wave motion or oscillations. In order to successfully complete the task, students had to

- hold the spring tension constant while varying the frequency, to create a graph of wavelength vs frequency,
- select a curve fit (often more than one curve fit matched the data), and interpret the values and the units of the curve fit parameters in a way that led to an algebraic expression,
- express the model in their own words,
- use their models to predict values that extended beyond their measurement region.

The model-breaking part of the assessment (Appendix B) was based on the Iron disk tossed onto cart video (Fig. 3) [43]. The video shows a very messy collision as the disk bounces and slides on the cart before coming to equilibrium. Students were provided with the model $p_i = p_f$, i.e., that the initial momentum is equal to the final momentum of a closed system. Then they were asked to evaluate whether the behavior of the disk and cart conformed or did not conform to that model. Since students often have the impression that momentum is only conserved in tidy collisions, we were curious how they would do analyzing this video. In order to successfully complete the task, students had to

- estimate the measurement uncertainty in the position of the disk and cart,

- use the (provided) uncertainty in the mass,
- combine the individual mass uncertainties to get the uncertainty in the combined mass,
- propagate the uncertainty in the displacement to get values for the uncertainty in the velocity and momentum,
- use the final estimate of the uncertainties to evaluate whether their measured p_i and p_f values agreed.

C. Rubrics

The model-making rubric awarded one point for each of the following:

M1. varying the frequency

M2. measuring the frequency

M3. measuring the wavelength

M4. making a graph of wavelength vs frequency

M5. fitting the graph to an inverse or power function

M6. writing the equation of the fit

M7. determining that wavelength is inversely proportional to frequency.

M8. predicting what will happen to the wavelength if the frequency is doubled

M9. identifying what physical quantity is represented by the curve fit parameters

M10. identifying the proper units for the curve fit parameters

A total of ten points were possible on this part of the assessment. Although most points were awarded as either full credit (one point) or no credit (zero points) partial points were awarded where appropriate, for example students earned ½ point on rubric element M5 if they *did* fit the graph to a curve but *did not* use an inverse fit (or a power curve with a coefficient of -1). Likewise, students earned ½ point on rubric element M7 if they indicated that there was an inverse relationship between wavelength and frequency without specifying that the relationship was inversely *proportional*.

The model-breaking rubric awarded one point for each of the following:

B1. writing down the uncertainty in the mass (which was provided in the instructions)

B2. propagating the absolute uncertainties in the mass of the disk and cart to determine the uncertainty in the combined disk and cart system after the collision

B3. estimating the uncertainty in the displacement of the disk and/or cart

B4. using the uncertainty in the displacement or time to determine the uncertainty in the velocity

B5. calculating the uncertainty in the momentum

B6. making any reference to absolute uncertainty, either by explicitly mentioning the term or by recording an absolute uncertainty value

B7. making any reference to relative uncertainty, either by explicitly mentioning the term or by recording a relative uncertainty value

B8. combining uncertainties using the square root of the sum of squares

B9. using uncertainties to answer the question of whether p_i is actually equal to p_f

A total of nine points were possible on this part of the assessment. Although most points were awarded as either full credit (1 point) or no credit (zero points) partial points were awarded for answers that were partially correct. For example, if students made an attempt to combine the uncertainties in the cart and disk masses but did so incorrectly (by simply adding the values rather than by finding the square root of the sum of their squares) they earned ½ point. Likewise, when students were asked to evaluate if the data matched the model they would earn a half of a point for mentioning uncertainty, but were required to refer to their specific uncertainty values to receive full credit.

The assessments were scored by three individuals. Two of the individuals were blinded to the treatment groups. The third individual was the instructor of one of the class sections and thus was not blinded to the students in his section, but was blinded for the other three sections. Interrater agreement was measured using Fleiss' kappa which indicated almost perfect agreement (0.88 and 0.90 for model making and model breaking, respectively) between scorers, where "agreement" was defined as two scores having the same value plus or minus 1. The average of the three scores for each rubric point was used to calculate the corresponding statistics.

Students were not provided the rubric. Rather, the mid-semester model-making and model-breaking activities were designed to lead students through the steps that would ultimately earn them points on the assessment rubric. The amount of scaffolding decreased throughout the semester so that the final assessments were minimally scaffolded.

D. Analysis

Comparisons were made to determine the effectiveness of each of the treatments. Because the data sets did not conform to a normal distribution as measured by Shapiro Wilk tests, the nonparametric two-tailed Mann-Whitney U test with a significance level of 0.05 was used to determine if groups differed by more than would be expected by chance. For all statistically significant results the effect size was calculated using the z score divided by the square root of the total number of samples, i.e., $z/\sqrt{N}$. An effect size of 0.1 is considered small. An effect size of 0.3 is considered medium, and an effect size of 0.5 is considered large.

V. RESULTS

A. Summary table

Table I provides a detailed summary of the data broken down by class section, treatment group, and evaluation metric

TABLE I. Detailed information about the performance of each treatment group within each class section on each of six measures. Please note that because many of the data sets were not normal all of the statistical comparisons in this study used the nonparametric Mann-Whitney U test, which is based on median values, and not *mean* values. The sample size, *n*, reflects either individuals (FMCE data) or small lab groups (non-FMCE data).

		Model-making treatment group			Model-breaking treatment group		
	Class section	Average	Standard deviation	***n***	Avgerage	Standard deviation	***n***
FMCE Pre-test scores	1	8.7	5.2	*16*	7.8	3.2	*19*
	2	7.9	2.9	*17*	9.6	8.9	*18*
	3	10.9	8.1	*15*	11.6	5.2	*10*
	4	14.5	14.9	*10*	9.2	5.7	*11*
	Total	**10.0**	8.1	*58*	**9.3**	6.2	*58*
FMCE Normal-ized gains	1	19.6	27.5	*16*	22.1	19.7	*19*
	2	20.6	22.7	*17*	20.9	29.2	*18*
	3	26.6	30.8	*15*	34.6	24.9	*10*
	4	23.0	29.3	*10*	22.4	27.3	*11*
	Total	**22.3**	26.9	*58*	**23.9**	25.1	*58*
Neutral quest. 1	1	0.64	0.24	*7*	0.61	0.49	*9*
	2	1.0	0.0	*6*	0.86	0.38	*7*
	3	0.75	0.42	*6*	0.5	0.41	*4*
	4	0.75	0.5	*4*	1	0	*4*
	Total	**0.78**	0.33	*23*	**0.73**	0.42	*24*
Neutral quest. 2	1	0.86	0.24	*7*	0.39	0.42	*9*
	2	0.58	0.49	*6*	0.57	0.45	*7*
	3	0.42	0.38	*6*	0.5	0.41	*4*
	4	0.75	0.5	*4*	0.5	0.41	*4*
	Total	**0.65**	0.41	*23*	**0.48**	0.4	*24*
Model-making Assess.	1	7.8	1.2	*7*	6.8	1.4	*9*
	2	8.6	1.3	*6*	7.0	0.9	*7*
	3	7.6	1.5	*6*	6.4	0.7	*4*
	4	8.9	1.3	*4*	6.4	1.8	*4*
	Total	**8.1**	2.1	*23*	**6.7**	1.8	*24*
Model-breaking Assess.	1	0.8	1.1	*7*	4.6	2.6	*9*
	2	1.5	2.7	*6*	3.6	3.3	*7*
	3	0.5	0.8	*6*	3.0	2.3	*4*
	4	0.3	0.7	*4*	5.1	3.5	*4*
	Total	**0.8**	1.5	*23*	**4.1**	2.9	*24*

B. Similarity of treatment groups

Analysis indicated no significant difference in FMCE pretest scores between the treatment groups, $z\ (113) = 0.17$, $p = 0.87$. There was also no significant difference between treatment groups in terms of FMCE pre and postnormalized gains, $z\ (109) = 0.37$, $p = 0.71$.

Analysis indicated no significant difference between treatment groups in the scores for either of the neutral questions:

- Finding the initial momentum $z(47) = 0.32$, $p = 0.75$
- Finding the final momentum $z(47) = 1.48$, $p = 0.14$

C. Model-making assessment

A model-making score was calculated for each small group out of ten possible points on the model-making assessment. There were 47 total groups. The model-making treatment group did significantly better on this part of the assessment than the other group, $z = 3.3$, $p = 0.0008$. For a total of $N = 47$ samples this corresponds to an effect size ($\mathrm{z}/\sqrt{N}$) of 0.49, or just shy of the 0.5 threshold for a large effect.

When the ten rubric points on the model-making assessment were analyzed individually, two of the points displayed a significant differentiation between the two treatment groups with the model-making treatment group outperforming the other group in both cases. The point awarded for determining that wavelength is inversely proportional to frequency (point M7 on the list above) produced a z score of 2.5 and a p value of 0.012. The size of the effect was 0.37. Likewise, the rubric point awarded for identifying what physical quantity is represented by the curve fit parameters (point M9 on the list above) displayed a significant difference between the treatment groups. The z score in this case was 2.6 and the p value was 0.0096. The effect size was 0.38.

D. Model-breaking assessment

A model-breaking score was calculated for each small group out of nine possible points on the model-breaking test. The model-breaking treatment group did significantly better on this part of the assessment than the other group, $z = 4.1$, $p =< 0.001$. For a total of $N = 47$ samples this corresponds to an effect size $(z/\sqrt{N})$ of 0.6 which signifies a large effect.

When the nine rubric points on the model-breaking assessment were analyzed individually, the model-breaking treatment group significantly outperformed the other group on all of the points. The size of the effect was at least 0.32 for all of the rubric points. It was greater than 0.5 for three of the points: B4, B6, and B8. Point B4 assessed whether students used the uncertainty in the displacement (or time) to determine the uncertainty in the velocity. Point B6 was awarded if students made a reference to absolute uncertainty in the process of determining if the observed events conformed to a given model. Students earned point B8 by combining uncertainties using the square root of the sum of squares.

VI. SUMMARY

The data analyzed in this article indicate that there were no significant differences between the treatment groups in terms of the FMCE prescore, FMCE normalized gains, or their performance on two neutral questions that were embedded in the final assessment. In addition, the results show that each treatment group significantly outperformed the other on the summative test that was designed to measure the knowledge and skills their treatment had emphasized. The groups that had completed five model-making activities throughout the semester did significantly better on the model-making assessment than the groups that had been in the other treatment. Likewise, the groups that had completed the model-breaking activities outperformed the other treatment on the model-breaking assessment.

VII. CONCLUSIONS AND DISCUSSION

Constructing accurate useful models often involves the cyclical, iterative process of proposing a model that works under some limited conditions and then testing that model under new conditions to see if it still applies. If the model no longer fits, then the previous model is refined or replaced and the process continues. This pattern can be seen in ISLE cycles (Investigative Science Learning Environments) [28], in Windschitl's "generation, testing, and revision of scientific models" [11], and in the iterative refinement described by Clement [37]. Two important parts of this process are developing a provisional model (i.e., model making) and then critically testing it in a new situation (i.e., model breaking).

Like most skills, being able to develop robust quantitative models takes practice and there are several indications that there is significant room for improvement within the status quo [44]. Researchers de Jong and van Joolingen have shown that students struggle at nearly every step in the process from generating hypothesis and designing experiments to implementing a systematic pan and interpreting the data [45]. Windschitl *et al.* indicate that even when instructors do incorporate inquiry activities into their classes the activities often miss the mark [11].

It is clear from the literature that the cyclical process of creating models can help students acquire scientific abilities [28,29], and that such a process can be implemented using apparatus-based labs and simulations [46]. What has not been known, until now, is if activities based on direct measurement videos could also help students acquire some of these skills. Overall, the results of this study indicate that DMV-based activities *can* produce significant learning outcomes for both of these skills. This is an encouraging result since many science instructors struggle with effectively fostering these skills. The fact that DMVs can produce robust learning of laboratory skills will give instructors more options to help their students acquire these skills. In addition, the fact that such learning can occur in a web-based environment rather than in an apparatus-based laboratory gives teachers more options about when and how to address these skills.

In the future, we would like to close the circle and measure students' ability to complete both steps sequentially: to develop a model under one set of circumstances and then be forced to refine it given a new set of conditions. For example, a student would discover the inverse squared nature of the Coulomb interaction between two charged spheres that are far apart and then discover a deviation from that ideal behavior as charge polarization within the spheres becomes apparent when the spheres are brought close to each other. Letting students see both aspects of models (their usefulness and their limits) is a worthy and important goal for science instructors that may help students establish the connection between what they are learning and the real world. While instructors expect their students to do that automatically, the research shows that students' ability to link their learning to the real world often *decreases* throughout their introductory physics courses [47,48].

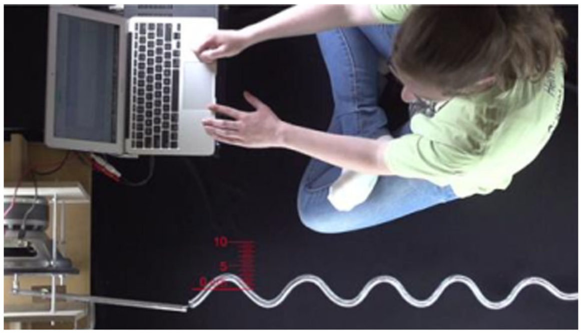
ACKNOWLEDGMENTS

This work is funded by the National Science Foundation (Grants No. 1245268 and No. 1644458), the American Institute of Physics, and the University of Wisconsin System. This work was produced under a grant from the UW-Extension SBIR Advance through its partnership with the WI Economic Development Corporation. The authors also acknowledge the important support of their home institutions: ISD No. 197, the University of Wisconsin River Falls, and the Science Education Resource Center.

APPENDIX A: MODEL-MAKING PART OF THE FINAL ASSESSMENT

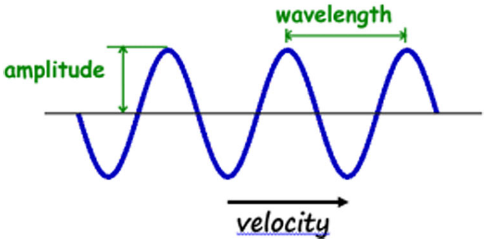
This assessment is intended to see how well you are able to use measurements, data, and calculations to determine the mathematical relationship between quantities. You are not expected to have prior knowledge about the physics of waves to be able to complete these questions.

Open up the Waves Properties video (near the bottom of this page: bit.ly/studentDMV)



For this activity you will need to know the following vocabulary:

- amplitude is the height of the wave measured from the center. It is often measured in m or cm.
- wavelength is the *distance* between two consecutive crests of the wave (measured in meters)
- period is the *time* between two consecutive crests. It is measured in seconds.
- frequency is the inverse of period (1/Period) It is measured in units of 1/seconds which has a special name. It is called Hertz, or Hz for short.
- velocity is how fast the wave is moving through space. It is measured in m/s.



(1) Use measurements from the videos to determine the mathematical relationship between the wavelength of a wave (in meters) and the frequency of that wave (in Hz). Write the resulting relationship below. Print off your graph with any analysis showing.

Note 1: Your performance will be judged largely on how well you show your process. In the space below please detail what you measured including which parameter was selected from the bottom toolbar (frequency, amplitude, and tension parameter), how you used your data, and what conclusions you were able to draw. Feel free to use an additional paper if you need more room to work.

Note 2: To improve the accuracy of your results please measure the period by timing over four cycles and then dividing by four. Remember to start counting cycles at zero (not one) when you zero the stopwatch.

(2) Describe the relationship between these two mathematical quantities. What does the shape of the graphed curve tell you about how these two quantities are related?

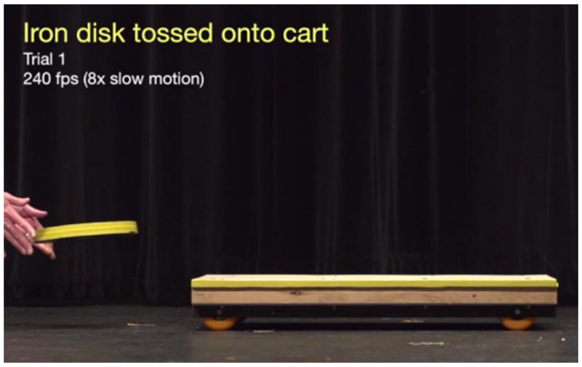
(3) How confident are you that your answer above is correct? (circle a number from 1 to 6 below) I'm certain that I'm correct 6 5 4 3 2 1 I am completely uncertain

(4) Based on the relationship that you found in your graph what would happen to the wavelength if you doubled the frequency?

(5) What are the units for your fit parameters? Describe the meaning of each of the parameters.

APPENDIX B: MODEL-MAKING PART OF THE FINAL ASSESSMENT

Go to bit.ly/studentDMV and open *Iron disk tossed onto cart* video under the Impulse & Momentum



For this lab you will need to know the following vocabulary:

- momentum is the velocity of the object times its mass. Momentum is a vector and is measured in (kg m/s). It is represented by the letter "p".
- closed system is a set of objects that aren't affected by things outside of the set.
- conservation of momentum is the idea that the total "p" of a closed system stays the same.

Conservation of momentum is a useful model that is often used to describe what happens when physical objects in a closed system interact. It can be summarized in the following equation $p_i = p_f$ which stands for: $\mathrm{momentum}_{\mathrm{initai}} = \mathrm{momentum}_{\mathrm{finai}}$.

This video depicts an event. Your job is to determine if the events in the video conform to the model (that $p_i = p_f$). The mass of the disk is 11.6 ± 0.01 kg & the cart mass is 23.4 ± 0.01 kg. Use the 240 fps setting.

(1) Considering only the horizontal motion of the iron disk and the cart, do the events in the video conform to the model? *Your performance will be judged on how well you support your answer. In the space below please detail your reasoning and use observations, measurements with uncertainties, and calculations to support your conclusions. Feel free to use an additional paper if you need more room to work.*

(2) How confident are you? (circle a number from 1 to 6 below) I'm certain that I'm correct 6 5 4 3 2 1 I am completely uncertain

(3) Assuming that the event did not conform to the model which of the following could be a possible explanation (circle all that apply). *NOTE: Please answer this* question even *if your results confirmed the model.*

(a) The floor is not level (if you chose this answer indicate which way it might be tilting).

(b) There is some friction between the iron disk and the cart.

(c) There is some friction between the cart and the stage

(d) The stated framerate of the video is incorrect (if you chose this answer indicate whether the actual frame rate must be higher or lower than the indicated frame rate).

[1] D. R. Dounas-Frazer, K. L. Van De Bogart, M. R. Stetzer, and H. J. Lewandowski, Investigating the role of model-based reasoning while troubleshooting an electric circuit, Phys. Rev. Phys. Educ. Res. **12**, 010137 (2016).

[2] D. F. Treagust, G. Chittleborough, and T. L. Mamiala, Students' understanding of the role of scientific models in learning science, Int. J. Sci. Educ. **24**, 357 (2002).

[3] *America's Lab Report: Investigation in High School Science*, edited by S. Singer and H. A. Schweingruber (The National Academies Press, Washington DC, 2005).

[4] D. Hestenes, Wherefore a science of teaching?, Phys. Teach. **17**, 235 (1979).

[5] D. Hestenes, Toward a modeling theory of physics instruction, Am. J. Phys. **55**, 440 (1987).

[6] M. Wells, D. Hestenes, and G. Swackhamer, A Modeling Method for high school physics instruction, Am. J. Phys. **63**, 606 (1995).

[7] http://modelinginstruction.org/.

[8] https://en.wikipedia.org/wiki/David_Hestenes.

[9] L. McDermott, P. Shaffer, and C Constantinou, Preparing teachers to teach physics and physical science by inquiry, Phys. Educ. **35**, 411 (2000).

[10] E. Brewe, Modeling theory applied: Modeling Instruction in introductory physics, Am. J. Phys. **76**, 1155 (2008).

[11] M. Windschitl, Jessica Thompson, Melissa Braaten, beyond the scientific method: model-based inquiry as a new paradigm of preference for school science investigations, Sci. Educ. **92**, 941 (2008).

[12] E. Etkina and S. Murthy, *Design labs: Students' expectations and reality 2005 Physics Education Research Conference* (2006).

[13] C. V. Schwarz, B. J. Reiser, E. A. Davis, L. Kenyon, A. Achér, D. Fortus, Y. Shwartz, B. Hug, and J. Krajcik, Developing a learning progression for scientific modeling: making scientific modeling accessible and meaningful for learners, J. Res. Sci. Teach. **46**, 632 (2009).

[14] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wiemann, New Instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[15] P. Bevington and K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences* (McGraw Hill, New York, 1969).

[16] S. Allie, A. Buffler, L. Kaunda, B. Campbell, and F. Lubben, First-year physics students' perceptions of the quality of experimental measurements, Int. J. Sci. Educ. **20**, 447 (1998).

[17] D. L. Deardorff, North Carolina State University, Ph.D. thesis, (2001).

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

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

[20] U.S. Dept. of Education, Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies (Revised 2010).

[21] http://www.vernier.com/products/books/pva/.

[22] This year as part of an NSF I-Corp award (No. 1644458) we interviewed over 100 science educators. When asked about their use of simulations a majority of teachers said

that while they did use simulations they worried that their students thought the simulations didn't reflect the real world.

[23] http://www.compadre.org/ivv/.

[24] P. W. Laws, M. C. Willis, D. P. Jackson, and K. Koenig, Using research-based interactive video vignettes to enhance out-of class learning in introductory physics, Phys. Teach. **53**, 114 (2015).

[25] http://www.compadre.org/ivv/research/PERoutcomes.cfm.

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

[27] D. Brookesand E. Etkina, Physical Phenomena in Real Time, Science **330**, 605 (2010).

[28] A. Karelina and E. Etkina, When and how do students engage in sense-making in a physics lab? *Proceedings of the Physics Education Research Conference 2006 Syracuse, NY* edited by L. McCullough, L. Hsu, and P. R. L. Heron (American Institute of Physics, College Park, MD, 2007).

[29] S. Murthy and E. Etkina Development of scientific abilities in a large class, *Proceedings of the Physics Education Research Conference 2004 Sacramento, CA* edited by J. Marx, P. R. L. Heron, and S. V. Franklin (American Institute of Physics, College Park, MD, 2005).

[30] http://Paer.rutgers.edu/resources.php.

[31] C. Wieman and N. G. Holmes, Measuring the impact of an instructional laboratory on the learning of introductory physics, Am. J. Phys. **83**, 972 (2015).

[32] C. Wieman Comparing Cognitive Task Analyses of Experimental Science and Instructional Laboratory Courses, Phys. Teach. **53**, 349 (2015).

[33] http://serc.carleton.edu/dmvideos/videos.html.

[34] http://serc.carleton.edu/dmvideos/players/keep_time.html?hide_banner=true.

[35] http://s3-us-west-2.amazonaws.com/dmvideos.org/players/waves_grid/waves_grid_3d.html.

[36] B. M. Zwickl, N. Finkelstein, and H. J. Lewandowski, Incorporating learning goals about modeling into an upper-division physics laboratory experiment, Am. J. Phys. **82**, 876 (2014).

[37] J. Clement, Model based learning as a key research area for science education, Int. J. Sci. Educ. **22**, 1041 (2000).

[38] It could be argued that performing an experiment and collecting data are separate skills from model making, i.e., that one could construct a model using extant data. We have decided to include them in our definition of model making since that process of going from a phenomenon to a model is so integral to the scientific method.

[39] https://www.aapt.org/Resources/upload/LabGuidlinesDocument_EBendorsed_nov10.pdf.

[40] N. G. Holmes and C. E. Wieman, Assessing modeling in the lab: Uncertainty and measurement, *BFY Proceedings*, edited by M. Eblen-Zayas, E. Behringer, and J. Kozminski (American Association of Physics Teachers, College Park, MD, 2015).

[41] http://Scaleup.ncsu.edu.

[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.13.020106 for summary of the five activities performed by the model-making group and the five activities performed by the model-breaking group.

[43] http://s3-us-west-2.amazonaws.com/dmvideos.org/players/disk_onto_cart/disk_onto_cart.html.

[44] The National Research Council's National Science Education Standards (1996).

[45] T. de Jong and W. R. van Joolingen, Scientific discovery learning with computer simulations of conceptual domains, Rev. Educ. Res. **68**, 179 (1998).

[46] N. D. Finkelstein, W. K. Adams, C. J. Keller, P. B. Kohl, K. K. Perkins, N. S. Podolefsky, and S. Reid When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment, Phys. Rev. ST Phys. Educ. Res. **1**, 010103 (2005).

[47] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student Expectations in Introductory PhysicsAm. J. Phys. **66**, 212 (1998).

[48] K. K. Perkins, M. M. Gratny, W. K. Adams, N. D. Finkelstein, and C. E. Wieman, Towards characterizing the relationship between students' self-reported interest in and their surveyed beliefs about physics, AIP Conf. Proc. **818**, 137 (2016).