Impact of Bayesian updating activities on student epistemologies

Aaron R. Warren*

Purdue University Northwest, Westville, Indiana 46391, USA

(Received 13 August 2019; published 3 January 2020)

The evaluation of hypotheses, and the ability to learn from critical reflection on experimental and theoretical tests of those hypotheses, is central to an authentic practice of physics. A large part of physics education therefore seeks to help students understand the significance of this kind of reflective practice and to develop the strategies required to accurately update their belief in the utility of various hypotheses and models. Prior work has introduced Bayesian updating activities as one potential means for cultivating such reflective practice within the context of introductory physics courses. These activities are not fixed pieces of curricular matter, but are better thought of as codified practices that are incorporated within lectures, labs, homework, and exams, and which are adaptable to a broad range of course formats. The Bayesian updating activities engage students in the use of hypothetico-deductive reasoning to test a hypothesis or model, followed by Bayesian updating at the conclusion of this test to update their subjective confidence in the hypothesis or model that was tested. Prior work has identified significant gains in pre- and post comparison of student scores on the Epistemological Beliefs Assessment for Physical Science (EBAPS) in introductory algebra-based courses. Here, we conduct a quasi-experimental study of the impact of Bayesian updating activities on student EBAPS scores in introductory calculus-based courses. Our analysis examines the impact to the overall EBAPS score, the subscores for each of the five original axes identified by the authors of the EBAPS, and the subscores for five alternative axes that were recently identified in other work via factor analysis of student responses [Johnson and Willoughby, Phys. Rev. Phys. Educ. Res. **14**, 010135 (2018)]. The results of our analysis show meaningful and credible gains on the overall EBAPS scores as well as for a multitude of the subscores. These gains are noteworthy due to their strength and their ability to be achieved with activities that were implemented as relatively minor alterations to a traditional course structure.

DOI: 10.1103/PhysRevPhysEducRes.16.010101

I. INTRODUCTION

A. Student epistemologies

Epistemological beliefs are the collective body of assertions that students ascribe to which influence subjective norms and attitudes regarding the doing and learning of physics. These beliefs may not be consciously articulated by students, and instead may be implicit and therefore crafted *ad novo* in response to questions or situations that are unlike anything the student has considered before. These beliefs, despite their perhaps initially tenuous nature, have been shown to significantly impact student performance in multiple ways, such as student achievements [1], conceptual comprehension [2], motivation [3], learning strategy [4], self-evaluation [5], and subject matter comprehension [6,7].

*arwarren@purdue.edu

In order to prepare for the contemporary and future workplace, physics students (including nonmajors) are generally expected to exit a class with greater proficiency in a wide range of scientific practices [8–12]. A significant portion of these practices necessarily involves epistemic reflection on data, hypotheses, and models in physics. Unfortunately, most students, even those in many nontraditional course formats, appear to complete their introductory physics courses having generally regressed in their understanding of the standards and practices of physics [13–16]. The design of a course can have some impact though, as gains on attitudinal assessments have been demonstrated in courses employing Physics by Inquiry [17], Physics of Everyday Thinking [18], Modeling Instruction [19–21], Investigative Science Learning Environment (ISLE) [22,23], and other designs [24,25].

One consistent feature shared by these approaches is the recurring elicitation of student evaluations of hypotheses and models. Evaluation strategies can generally be framed within the hypothetico-deductive (HD) process [26–28]. The HD process serves the important function of eliciting and structuring a coherent body of reasoning that allows the

TABLE I. A schematic outline of the hypothetico-deductive process.

IF	[the hypothesis to be tested is assumed true]
AND	[a test is planned under certain assumed conditions]
THEN	[a prediction is deduced]
AND/BUT	[results of the test, including associated uncertainties, were obtained]
THEREFORE	[estimated likelihood of hypothesis should/should not be changed]

learner's confidence in the hypothesis to be modified in proportion to the epistemic power of the evaluative judgment that is rendered. The HD process is outlined in Table I, and stems from considerations of the reasoning patterns that generate changes in the epistemic status of scientific hypotheses and models [29].

Consistency may be important here in terms of affording the students a single cognitive structure that is gradually made more robust and productive by repeated use across a range of subject matter topics. The repetition of use may afford a greater degree of cognitive chunking, so that students retain greater cognitive capacity to attend to more details simultaneously and craft a more robust and expert-like belief. For example, ISLE presents students with a model of scientific activity called the ISLE cycle and makes repeated and consistent use of this cycle to structure student activities in lecture, recitation, and labs.

While these curricula demonstrate strong student gains on student epistemological beliefs as well as a variety of other positive outcomes, they often appear to have a relatively high perceived barrier to entry among physics instructors. A primary factor is reported to be instructor concerns about the requisite time to adopt and employ such curricula [30,31]. Thus, there is strong motivation to develop and test materials that are easier to deploy by instructors who prefer to utilize a traditional pedagogical approach, and yet which can positively and significantly impact student epistemological beliefs.

The Bayesian updating activities are intended as a step in that direction. They are distinct in that they also engage students in the evaluation of hypotheses and models, but intend to do so in a low-profile "plug-and-play" fashion to facilitate their easy adoption in nearly any course format an instructor may favor, whether traditional or otherwise. Ideally, these activities would enable significant gains to be made in student epistemological beliefs at a minimal cost of class time and instructor preparation. Bayesian updating activities are designed to engage students in evaluations of hypotheses and models that are consistently structured by the HD process, and then engage students in epistemic reflections that are consistently structured by the use of Bayes's theorem.

B. Bayesian updating activities

Depending on the nature of the hypothesis being tested, there are two broad types of Bayesian updating activities. If the hypothesis being evaluated makes assertions about data, such as the presence or absence of patterns in the data, then we call it a *direct evaluation*. If the tested hypothesis makes assertions about the relationships between physical models (such as limit cases) or between a physical model and a general theoretical principle (such as energy conservation), we call it an *indirect evaluation*. The term "indirect" is used for the latter category because prior data have already been used in the establishment of some models or principles, and those data are therefore being indirectly leveraged to test a hypothesis concerning relations between those models or principles. This type of evaluation is generally less familiar to students, as physics is typically the first course where they have exposure to the idea that one can test hypotheses and models using thought experiments that leverage prior data instead of having to do a physical experiment to collect new data.

After completing the HD process and deciding whether one's confidence in the hypothesis should increase or decrease (or remain unchanged), the actual updating of the learner's confidence level is done using Bayes's theorem. For our purposes, this is formulated as

$$P(H|E) = \frac{P(H) * R}{P(H) * R + 1 - P(H)}, \tag{1}$$

where $P(H|E)$ is the probability of hypothesis H being true given the newly acquired evidence E, $P(H)$ is the initial probability of hypothesis H being true before consideration of the new evidence, and

$$R = \frac{P(E|H)}{P(E|\neg H)} \tag{2}$$

is a Bayes factor, with $P(E|H)$ the probability of the evidence E being produced if the hypothesis H was true, and $P(E|\neg H)$ the probability of the evidence E being produced if the hypothesis H was false. Thus, this Bayes factor is the ratio by which a particular piece of evidence is relatively more likely to be produced by hypothesis H than by its converse, $\neg H$, and thus encodes the inferential power of the evidence with regards to H.

For example, a strong confirmatory result means that the evidence can be much better explained by the hypothesis than otherwise, and the student who wishes to update their confidence in the hypothesis should choose a large R in such cases (i.e., $R \gg 1$). A strong disconfirmatory result means that the evidence is much better explained by assuming the

hypothesis is false than otherwise, and therefore the students should choose a small R ($0 < R < 1$). A null result is when the evidence offers no discrimination between H and $\neg H$, in which case the student should select $R = 1$.

In general, the choice of R is subjective to the extent that there is no single value that "should" be chosen because of the unavoidable presence of experimental errors that make it impossible to definitively say, for example, whether an experiment result is truly null, or very weakly confirmatory, or very weakly disconfirmatory. However, the choice a student makes for R is constrained to the degree that it must be defensible based on reasoned consideration of the quantitative results and of any possible sources of error that were not quantified during the experiment. Guidelines for the estimation of R are provided in Table II.

This approach can be used for both direct evaluations (such as in laboratory reports) and indirect evaluations (such as in lecture and homework questions). Activities for each case were first introduced in Ref. [33]. Direct evaluation activities are completed during the laboratories, being embedded as part of the lab report writing process. The intent of these activities is to evoke and to structure student reflection on the epistemic significance of testing experiments conducted during the labs. In particular, the goal of a testing experiment is framed as the acquisition and application of evidence to modify subjective confidence in a model or hypothesis. The importance of error analysis is consequently enhanced by having students consider whether and how specific errors may have produced false positives or negatives, and ultimately whether the results are confirmatory, null, or disconfirmatory. A summary section at the end of the report asks students to briefly encapsulate the logical flow and results of the experiment using the hypothetico-deductive structure, and to update their confidence in the model or hypothesis being tested by selecting and justifying a particular value for the updating coefficient R, then using Bayes's theorem to calculate their updated confidence. An example lab report guideline given to students is shown in Fig. 1. The direct evaluation activity is the summary section at the end.

TABLE II. Guidelines for estimation of R (adapted from Ref. [32]).

R	Interpretation
$< \frac{1}{150}$	$\neg H$ very strongly favored
$\frac{1}{150}$ to $\frac{1}{20}$	$\neg H$ strongly favored
$\frac{1}{20}$ to $\frac{1}{3}$	$\neg H$ substantially favored
$\frac{1}{3}$ to 1	$\neg H$ barely favored
1 to 3	H barely favored
3 to 20	H substantially favored
20 to 150	H strongly favored
>150	H very strongly favored

Indirect evaluation activities are completed during each class period as group work, then turned in to the instructor, and written feedback is provided by the instructor at the following class period. These activities typically ask students to first solve a standard end-of-chapter exercise. After creating their proposed solution, students state their confidence that their solution is correct and follow the hypothetico-deductive reasoning template to conduct a thought experiment (such as a special-case analysis) in order to test their solution. They then choose and justify a value for the updating coefficient R and use Bayes's theorem to update their confidence in their solution. The pedagogical motivation for these activities is to provide students with a consistent structure and clear motivation for reflecting and assessing their own work, and to guide them towards independent learning. An example in-class activity

Lab Report

Your lab report should follow this format, including a cover sheet with all names, the date, and a title at the beginning.

a. *Hypothesis:* Describe the hypothesis being tested and state your initial confidence level in the hypothesis. (1 pt)

b. *Experiment Design:* Describe (both verbally and with pictures) how you collected your data. Include information about steps taken to minimize systematic and random errors. (1 pt)

c. *Prediction:* What are you going to compare? If the hypothesis is correct, what should be your results? (1 pt)

d. *Data:* Include your Excel workbook with all data. (1 pt)

e. *Analysis:* Include all analyses of the data, including calculations, graphs, and statistical work, in your Excel workbook. In your report, state all equations used in your calculations, including the equations you derive for error propagation. (1 pt)

f. *Result:* Report your results, including the best estimate, total uncertainty, and 95% confidence interval for the frequency. State whether your result is consistent with the prediction you made in section C above. (1 pt)

g. *Error Analysis:* Identify and analyze sources of systematic error in your experiment that were not accounted for in your calculation of the total uncertainty of the frequency. For each source of error that you list, briefly describe why you identified it as a source of error, the direction it is likely to skew the results, and the estimated magnitude of the skew. (1 pt)

h. *Evaluate Results:* Make a judgment whether your experiment produces a confirmatory, null, or disconfirmatory result. Justify your judgment on the basis of your prediction, results, and error analysis. (2 pts)

i. *Summary:* Summarize the logical structure of this experiment using the IF… AND… THEN… AND/BUT… THEREFORE… structure. Choose a value for the updating coefficient R, and justify that choice. Update your confidence in the hypothesis using Bayes' Theorem. (1 pt)

FIG. 1. An example lab report guideline for the E condition, taken from the final lab, which asks students to test a hypothesis regarding the frequency of an oscillator by measuring the tension required to achieve different standing wave modes. The direct evaluation activity is embedded as the summary section, part (i). A total of 10 lab reports, each including embedded direct evaluation activities in the form of summary sections to be written by students, were assigned in the E sections during the semester.

Activity 4 **Names:**

1. Exercise 3.22

2. Exercise 3.21

3. **a.** What is your initial confidence in your solution to Part A of Exercise 3.21?

b. Do a thought experiment to test your solution using the IF... AND... THEN... AND/BUT... THEREFORE... process.

c. Update your confidence using Bayes' Theorem. Justify your choice of the updating coefficient *R*.

Updating Coefficient *R*:

R	Interpretation
< (1/150)	Very strong disconfirmation
(1/150) to (1/20)	Strong disconfirmation
(1/20) to (1/3)	Substantial disconfirmation
(1/3) to 1	Weak disconfirmation
1	Null
1 to 3	Weak confirmation
3 to 20	Substantial confirmation
20 to 150	Strong confirmation
> 150	Very strong confirmation

Bayes' Theorem:

$$C_f = \frac{C_i R}{C_i R + 1 - C_i}$$

d. Could your confidence ever (after finitely many thought experiments) reach 100% or 0%? Explain why or why not.

4. Exercise 3.41

FIG. 2. An example in-class activity from the *E* sections. Students must show all work, and work is graded only for completion, not correctness. The indirect evaluation activity item is No. 3 on this assignment. Student work was given formative written feedback at the beginning of the following class period for every in-class activity. During the semester, a total of 18 indirect evaluation activities were given to students in the *E* sections.

is shown in Fig. 2, where the indirect evaluation activity is No. 3.

Students are also given resource materials including guideline documents for thought experiments (which constitute the indirect evaluation activities) and for writing lab reports (which constitute the direct evaluation activity). Copies of these documents, as well as an introductory lab activity to familiarize students with the direct and indirect Bayesian updating activities, the answer key to this activity, an example lab manual, and a sample lab report given to students after turning in their first lab at the beginning of the third week of the semester) are all available in the Supplemental Material [34].

Compared to ISLE, Modeling, PBI, or other research-based curricula, the Bayesian updating activities have a lower implementation profile, potentially serving more as an add on to a traditional physics course than as a deeply integrated reform. That is not to say that Bayesian updating activities are in any way at odds with research-based curricular approaches, and in fact an initial study presented in Ref. [33] found early evidence of some significant positive gains on the Epistemological Beliefs Assessment for Physical Science (EBAPS) [35] due to the use of Bayesian updating activities in an introductory algebra-based course which was already ISLE-like.

It may be, however, that the relative gains produced by use of the Bayesian updating activities in that study relied on the fact that the course curriculum already included ISLE-like pedagogical practices which valued and engaged students in epistemological reflection and growth. Without that alignment, it is possible that the Bayesian updating activities would appear to be spurious or purposeless to students who are otherwise engaged in a traditional curriculum, undercutting their efficacy. This is a major concern, and one which motivates a study to address it by determining the relative epistemological gains produced by the addition of Bayesian updating activities to a traditional course. For reasons of logistics (primarily sample size) we have chosen in this study to examine their impact in a traditional calculus-based physics course, as opposed to algebra based.

According to the literature as summarized at the beginning of Sec. I, epistemological gains are generally expected to differ for students in a traditional course versus an ISLE-like course. Students in a traditional course would likely show losses on EBAPS while the ISLE-like course would likely produce gains. Thus, while the study in Ref. [33] indicated that Bayesian updating activities produced relative significant gains only on the overall EBAPS score and one subscale axis (Axis 2: Nature of Knowing and Learning), if the Bayesian activities are able to be effective on their own, we would expect much broader and more significant gains to be made relative to a traditional curriculum, because the ISLE-like curriculum was already addressing some of the same epistemological issues as the Bayesian updating activities. For example, ISLE already includes an array of evaluation activities as well as laboratory designs that engage students in epistemic reflection. While the precise nature of those activities in ISLE differs from the Bayesian updating activities, they are similar enough so that the added value of the Bayesian updating activities may be marginalized.

Considering the characteristics of each of the five axes of the EBAPS, one may expect gains on all of them to be produced by the use of Bayesian updating activities when employed in an otherwise traditional course. The indirect evaluations engage students in special-case and limit-case analyses that emphasize the coherent structure of physics knowledge (Axis 1: Structure of Scientific Knowledge). They are also explicitly constructionist as students craft arguments to critically evaluate their own work and modify their own beliefs (Axis 2: Nature of Knowing and Learning, Axis 5: Source of Ability to Learn). Both the direct and indirect evaluation activities aim to engage students in authentic expertlike revision of beliefs about hypotheses and models (Axis 4: Evolving Knowledge). The direct evaluation activities emphasize that hypotheses and models in physics are general constructs, applicable to a range of situations and phenomena, and intend to strengthen the connection between abstract

models and concrete physical situations (Axis 3: Real-life applicability).

II. RESEARCH QUESTION

This study aims to determine whether the Bayesian updating activities are able to produce significant epistemological gains in an otherwise traditional course. If so, this would indicate that the activities are indeed a low-cost, low-profile approach that can be used by instructors who wish to positively impact student epistemological beliefs, but who hesitate to adopt more extensive curricular reforms. Conversely, if the activities are found not to produce significant gains on all (or nearly all) of the EBAPS axes, it would indicate that their effectiveness is limited and perhaps reliant on integration with sufficiently aligned and more extensive curricular approaches such as ISLE.

III. METHODS

A. Design

We use a quasiexperimental design at a medium-sized regional campus of a land-grant institution. Students from one calculus-based introductory physics course on mechanics serve as the control group, and students from two other sections of the same course taught in different semesters by another instructor serve as the experimental group. Throughout this paper, the control section is labeled C while the two experimental sections are labeled $E1$ and $E2$ (or simply E when referring to the combination of the two). The total number of registered students in each section is $N_C = 55$, $N_{E1} = 34$, and $N_{E2} = 25$, with nearly all students in each section (≈90%) being male engineering majors in their first year of study. Roughly 60% of students in each section are first-generation college students. All sections of the course use the same textbook [36], similar laboratory experiments, similar in-class group-work activities, and similar online homework assignments composed of end-of-chapter items from the textbook. The principle difference between the two conditions was the use of Bayesian updating activities in the experimental sections. Both instructors are young Caucasian males who received strong student evaluation scores at the end of the course (scores of 4.5 to 4.9 out of 5.0) indicating comparable levels of student affect toward the courses and instructors.

The utilization of Bayesian updating activities was consistent across the $E1$ and $E2$ sections, utilizing activities such as those presented in Ref. [33]. There were a total of 11 lab periods, each 110 minutes long. One hour of the first lab period was used to introduce students to the HD process and Bayesian updating, and to model their use in both laboratory experiments (i.e., direct evaluation) and thought experiments relating to homework or exam questions (i.e., indirect evaluation). For pedagogical reasons, Bayes's theorem is presented to the students as

$$C_f = \frac{C_i * R}{C_i * R + 1 - C_i}, \tag{3}$$

where $C_i = P(H)$ and $C_f = P(H|E)$ are interpreted as a student's initial and final (updated) confidence in the hypothesis H. Students then spent the remainder of that lab period completing an activity that involved Bayesian updating for both direct and indirect evaluations. This activity also walks students through an example of repeated testing of a hypothesis, showing that regardless of the specific value for R that is chosen, the confidence any person has in a hypothesis will ultimately converge toward 0 or 1 so long as all parties qualitatively agree on whether the tests produce confirmatory or disconfirmatory results. The effect of different choices for R is only to change the rate at which a person's confidence in the hypothesis will asymptotically approach 0 or 1. The remaining 10 labs required students to work in groups of 2–3 to conduct experiments designed to directly evaluate particular hypotheses, and to write lab reports. The format of the labs is traditional in that students are given instructions on how to use the materials. Lab reports in the E condition were structured as direct evaluation activities, such as shown in Fig. 1. Lab report guidelines in the C condition were generally isomorphic except for the Bayesian updating component.

There were a total of 26 lecture periods during the semester for each section, with each period being 110 minutes long. Lectures throughout the semester included instructor modeling and student engagement in the HD process and Bayesian updating for indirect evaluations, with a total of roughly 2–3 hours (out of 48 total) spent this way. Student engagement was done via 18 in-class indirect evaluation activities that asked them to solve a standard end-of-chapter problem, and then to do an indirect evaluation of their solution in order to update their confidence in their solution's accuracy, as shown above in Fig. 2. The end-of-chapter problems were always quantitative exercises, and were typically of intermediate difficulty according to the textbook's ranking. Students worked in groups of 2–3 on the activities, which were turned in to the instructor at the end of each class period for attendance credit, and returned the following class period with comments. Comments emphasized the purpose and structure of the general HD reasoning pattern and Bayesian updating process as needed. For example, some students initially thought the purpose of these activities was to describe their reasoning for their solution instead of testing their solution and determining whether they should be more or less confident in the accuracy of their solution. Grades for these activities were not based on how well students did, credit was uniformly given to students who turned in any work and all instructor feedback was entirely formative. During the first half of the semester students gradually acclimate to these activities, and occasional whole-group discussions were held (lasting 5–15 minutes) discussing the purpose and execution of these

indirect evaluation activities. Weekly homework assignments included 1–2 such items that were graded based on the quality of student work. The homework was done online using the Pearson Mastering Physics platform, and these indirect evaluation items were created and added to the online assignments. Exams included one open-response question that required students to similarly evaluate their proposed solution to a particular exam question by conducting a thought experiment.

B. Instrument

Data were collected via administration of the EBAPS during the first week of classes and again during the final week. This instrument is a 30-item forced-choice assessment of students' personal epistemologies. Each item is scored on a scale from 0 to 4, with 0 representing a novicelike perspective and 4 representing an expertlike perspective. An overall score is then calculated by averaging the individual item scores and multiplying by 25 to produce a scale from 0 to 100. In addition, five subscores are calculated to place each student along five nonorthogonal axes of epistemic belief, each with values ranging from 0–100. The names of these axes are listed in Table III.

The design of the EBAPS is rooted in a theoretical framework of finely grained cognitive resources (comparable to diSessa's *p* prims [38]) which are thought to be contextually triggered. The attitudes expressed by students item-by-item may therefore vary due to contextual differences that influence the application of these resources. Students may not have given much thought to scenarios or questions at all like those posed by EBAPS, and thus their responses may not represent stable, fundamental beliefs. The intent of the EBAPS is to probe the epistemological stances students take, even though students may not knowingly have a stance on abstract epistemological issues such as coherent knowledge versus knowledge in pieces. The items grouped on a single axis are intended to pertain to a specific abstract epistemological issue that experts see as being implicit within each of those items.

TABLE III. EBAPS axis numbers and titles. The first five are the original axes from Ref. [35], and the second five (denoted with *) from Ref. [37].

Axis 1	Structure of scientific knowledge
Axis 2	Nature of knowing and learning
Axis 3	Real-life applicability
Axis 4	Evolving knowledge
Axis 5	Source of ability to learn
Axis 1*	Structure of science
Axis 2*	Innate ability vs hard work
Axis 3*	Source of ability to learn
Axis 4*	Nature of knowing and learning
Axis 5*	Quick learning

Recently, an exploratory factor analysis of student data has identified an alternative set of five dimensions within the EBAPS [37], which are also listed in Table III. These appear similar in some ways to the original five axes, with the major difference that the real-life applicability axis (axis 3) was insignificant in the recorded student response patterns. It should be noted that the data used in their factor analysis was drawn from introductory astronomy courses, and thus it is not known whether these same factors would be identified in a factor analysis of responses from introductory physics courses. For completeness, though, we think it wise to also examine student performance along this alternative set of axes, which we distinguish from the original five axes with an asterisk (*).

C. Procedure

All student responses were scored and included in our analysis, with pre- and post-test response rates of 93% and 78% for section *C*, 94% and 91% for section *E*1, and 96% and 88% for section *E*2. The relatively lower post-test completion for the *C* section is due to fewer students attending the final lab of the semester. This itself may stem from differences in class sizes, student populations, instructors, or even the Bayesian updating activities themselves which may have improved students' perceptions of the importance of the labs.

There were no identifiers given on student responses, so neither listwise deletion nor multiple imputation are possible. Instead, we conduct Bayesian MCMC estimation of the score distributions for each section (for both pre- and post-tests, separately). Our approach is based on the Bayesian estimation supersedes the *t*-test (BEST) method [39]. This generates an explicit distribution of credible values for parameter estimation of a distribution and uses those distributions to test for differences between groups. This approach is much more robust and informative than traditional distribution characterization and tests for distributional differences such as *t* tests. Default broad priors are used for all parameter estimates. Comparisons of the MCMC-generated pre- and post-test parameter distributions produce estimates of the gains made by each section (*C*, *E*1, *E*2) as well as for the combined *E* sections.

The parameter estimates for the pre- and post-test responses for the *E* and *C* conditions are then used to estimate the difference in gains between the two conditions on the overall EBAPS score, each original axis score, and each alternative axis score. In particular, the MCMC chains for the means that were produced during parameter estimation are differenced and a kernel density function is calculated to obtain a distribution of the credible differences of means for each score type (overall, axis 1, axis 2, etc.). The median and 95% highest density interval (HDI) of these distributions are then calculated and reported as the difference in gains made between the *E* and *C* conditions. A similar procedure is used to generate a

distribution of credible values for effect sizes by using the MCMC chains for the standard deviations of each distribution produced during parameter estimation. The R statistical computing language [40] was used for all analyses and employed the Bayesian MCMC sampling program JAGS [41], with scripts we have adapted from Ref. [39]. Our modified analysis scripts and instructions on their use are available from the author.

IV. RESULTS

A. Individual sections

Best estimates of the mean (and associated 95% HDIs) for each section's respective pre- and post-test results, and estimates of the differences in means, are listed in Tables IV–VI. If we wish to compare the combined performance of the E sections with the C section, we must first compare the pre- and post-test performances of the two E sections to determine if they are sufficiently similar as to justify their aggregation. Comparison of the pretest scores for the two experimental sections ($E1$ and $E2$) show no credible differences at the 95% level, indicating an acceptable level of homogeneity between the two groups on the pretest.

Considering the post-test scores and the gains made there is again general consistency between sections $E1$ and $E2$, although we see an interesting feature on axis 2*. Section $E2$ shows negative gains on this axis, of weak effect size and fairly low credibility, but comparing it with the gains from section $E1$ yields a more credible difference (likelihood 93.0%). We posit that if this difference is real it may be due to chance variations between the students in the two E sections. Overall, however, we feel it is reasonable to group sections $E1$ and $E2$ together as they show credibly similar pre- and post-test performance on the overall score and axes except for this one.

Another challenge in our quasiexperimental study is to determine whether the E and C conditions are initially similar enough as to allow differences in their post-test performances (and all derived quantities) to have convincing inferential power regarding the efficacy of the Bayesian updating activities. Examining the pretest scores, we note that there is broad agreement for section C and the two E sections, with the caveat that the C section shows credible differences (at the 95% level) on the Overall score as well as axis 2, axis 3, and axis 4*. For the moment, despite these differences, we choose to operate under the assumption that the populations of the three sections are initially uniform enough to produce reasonably valid conclusions of treatment efficacy based on comparisons of the post-test results. After examining the overall differences between the E and C conditions, though, we will return to this issue and

TABLE IV. EBAPS pre- and post-test scores for section C, the mean gains, and effect sizes. The best estimate for each quantity obtained via Bayesian MCMC is stated, along with the bounds of the associated 95% highest density intervals (HDIs).

Section C	Pre	Post	Mean gain	Effect size
Overall	$64.5^{67.1}_{61.7}$	$61.1^{64.6}_{57.7}$	$-3.4^{1.0}_{-7.7}$	$-0.305^{0.101}_{-0.751}$
Axis 1	$58.3^{61.9}_{54.8}$	$56.9^{61.4}_{52.9}$	$-1.4^{4.1}_{-6.9}$	$-0.112^{0.308}_{-0.547}$
Axis 2	$59.5^{62.8}_{56.2}$	$58.1^{61.5}_{54.7}$	$-1.4^{3.2}_{-6.2}$	$-0.130^{0.319}_{-0.535}$
Axis 3	$69.3^{74.6}_{64.2}$	$67.6^{74.1}_{60.9}$	$-1.8^{6.7}_{-10.2}$	$-0.089^{0.345}_{-0.508}$
Axis 4	$64.3^{69.5}_{58.6}$	$60.0^{67.0}_{53.3}$	$-4.4^{4.5}_{-13.3}$	$-0.212^{0.205}_{-0.652}$
Axis 5	$76.1^{81.7}_{70.8}$	$69.8^{76.5}_{63.3}$	$-6.3^{2.2}_{-14.6}$	$-0.310^{0.122}_{-0.715}$
Axis 1*	$75.3^{80.8}_{69.9}$	$67.2^{75.5}_{59.7}$	$-8.2^{1.8}_{-17.2}$	$-0.371^{0.042}_{-0.816}$
Axis 2*	$80.1^{86.2}_{74.3}$	$69.4^{76.9}_{62.0}$	$-10.7^{-1.2}_{-20.3}$	$-0.483^{-0.034}_{-0.899}$
Axis 3*	$72.2^{78.2}_{66.8}$	$69.8^{77.0}_{63.1}$	$-2.4^{6.5}_{-10.9}$	$-0.125^{0.328}_{-0.568}$
Axis 4*	$70.3^{77.3}_{63.3}$	$68.0^{76.9}_{58.9}$	$-2.3^{9.2}_{-13.5}$	$-0.087^{0.324}_{-0.514}$
Axis 5*	$67.7^{76.9}_{50.3}$	$62.9^{72.5}_{53.8}$	$-4.8^{6.9}_{-17.1}$	$-0.170^{0.245}_{-0.595}$

TABLE V. Best estimates and 95% HDIs for the pre- and post-test scores for section $E1$, and the mean gains.

Section $E1$	Pre	Post	Mean gain
Overall	$67.5^{70.4}_{64.3}$	$72.5^{75.7}_{69.3}$	$5.0^{9.4}_{0.4}$
Axis 1	$60.1^{63.9}_{56.5}$	$68.5^{73.0}_{64.0}$	$8.4^{14.0}_{2.2}$
Axis 2	$63.0^{68.2}_{58.0}$	$65.7^{70.5}_{61.1}$	$2.6^{9.3}_{-4.5}$
Axis 3	$78.8^{85.1}_{73.0}$	$85.4^{90.1}_{80.5}$	$6.6^{14.3}_{-1.2}$
Axis 4	$69.4^{75.9}_{62.9}$	$70.9^{76.5}_{65.4}$	$1.4^{9.8}_{-7.2}$
Axis 5	$78.1^{84.7}_{71.5}$	$83.7^{89.3}_{77.7}$	$5.6^{14.3}_{-2.9}$
Axis 1*	$76.3^{83.2}_{68.9}$	$84.7^{90.2}_{78.3}$	$8.4^{17.3}_{-0.6}$
Axis 2*	$80.8^{89.1}_{72.3}$	$86.6^{92.4}_{80.5}$	$5.8^{15.9}_{-4.2}$
Axis 3*	$75.9^{82.2}_{69.7}$	$79.2^{85.8}_{72.2}$	$3.3^{11.8}_{-6.3}$
Axis 4*	$77.0^{58.1}_{69.1}$	$84.4^{90.8}_{78.2}$	$7.4^{17.1}_{-2.9}$
Axis 5*	$71.1^{78.4}_{63.7}$	$73.2^{82.3}_{63.4}$	$2.0^{13.5}_{-9.8}$

TABLE VI. Best estimates and 95% HDIs for the pre- and post-test scores for section $E2$, and the mean gains.

Section $E2$	Pre	Post	Mean gain
Overall	$70.2^{73.3}_{67.1}$	$74.8^{79.2}_{70.2}$	$4.6^{9.9}_{-1.0}$
Axis 1	$62.8^{68.0}_{58.1}$	$71.2^{77.6}_{65.1}$	$8.4^{16.1}_{0.0}$
Axis 2	$67.1^{71.9}_{62.6}$	$71.4^{76.5}_{66.6}$	$4.3^{11.4}_{-2.2}$
Axis 3	$75.2^{83.0}_{67.4}$	$83.5^{92.6}_{74.7}$	$8.4^{20.4}_{-3.4}$
Axis 4	$70.6^{79.1}_{62.1}$	$71.9^{81.8}_{62.2}$	$1.3^{14.4}_{-11.5}$
Axis 5	$80.5^{81.8}_{73.3}$	$82.9^{90.5}_{75.2}$	$2.5^{13.2}_{-7.7}$
Axis 1*	$86.1^{91.1}_{80.5}$	$86.9^{95.2}_{77.6}$	$0.8^{9.9}_{-9.7}$
Axis 2*	$87.7^{94.4}_{80.4}$	$82.3^{91.1}_{73.9}$	$-5.5^{5.6}_{-16.6}$
Axis 3*	$79.4^{85.5}_{73.2}$	$85.9^{92.0}_{79.8}$	$6.5^{15.2}_{-2.0}$
Axis 4*	$81.7^{90.6}_{72.3}$	$81.4^{97.0}_{85.5}$	$9.8^{20.1}_{-1.1}$
Axis 5*	$72.1^{79.9}_{64.4}$	$77.6^{86.4}_{68.6}$	$5.5^{17.4}_{-6.0}$

TABLE VII. Best estimates and 95% HDIs for the pre- and post-test scores for the combined *E* sections, the mean gains, and the effect sizes.

E Condition	Pre	Post	Mean gain	Effect size
Overall	$68.6^{70.7}_{66.6}$	$73.5^{76.1}_{71.0}$	$4.9^{8.1}_{1.6}$	$0.586^{0.995}_{0.183}$
Axis 1	$61.2^{64.3}_{58.5}$	$69.6^{73.2}_{66.0}$	$8.4^{13.2}_{3.8}$	$0.722^{1.15}_{0.319}$
Axis 2	$64.8^{68.2}_{61.4}$	$68.0^{71.6}_{64.8}$	$3.2^{7.8}_{-1.7}$	$0.264^{0.637}_{-0.137}$
Axis 3	$77.2^{82.0}_{72.5}$	$84.7^{89.1}_{80.4}$	$7.4^{13.8}_{1.4}$	$0.468^{0.876}_{0.079}$
Axis 4	$69.9^{74.6}_{65.1}$	$83.4^{88.0}_{78.9}$	$1.4^{8.6}_{-8.3}$	$0.079^{0.472}_{-0.307}$
Axis 5	$79.2^{84.1}_{74.7}$	$83.4^{88.0}_{78.9}$	$4.3^{10.5}_{-1.9}$	$0.270^{0.682}_{-0.113}$
Axis 1*	$81.3^{86.0}_{76.5}$	$86.1^{90.8}_{81.1}$	$4.8^{11.2}_{-1.6}$	$0.346^{0.816}_{-0.128}$
Axis 2*	$83.8^{89.0}_{78.1}$	$84.7^{89.3}_{79.4}$	$0.8^{8.0}_{-6.1}$	$0.045^{0.436}_{-0.345}$
Axis 3*	$78.0^{82.0}_{73.9}$	$83.1^{87.5}_{78.9}$	$5.1^{10.8}_{-0.9}$	$0.407^{0.871}_{-0.084}$
Axis 4*	$78.6^{84.2}_{72.8}$	$87.0^{91.5}_{82.7}$	$8.4^{15.8}_{1.7}$	$0.473^{0.869}_{0.068}$
Axis 5*	$71.5^{76.9}_{66.7}$	$75.1^{81.3}_{68.8}$	$3.5^{11.6}_{-4.4}$	$0.174^{0.564}_{-0.218}$

consider the possibility of some bias relating to these pretest differences.

B. Overall *E* vs *C* conditions

We first note from Table IV that the *C* condition shows slightly negative epistemic gains, although these are mostly of weak effect and have low credibility. Exceptions to this include the more pronounced negative gains shown in the overall, axis 5, axis 1*, and axis 2* scores. This is generally consistent with the literature, which shows similarly negative epistemic impacts in traditional introductory physics courses. In contrast, the combined results of the *E* condition, summarized in Table VII and Figs. 3–5, show uniform epistemic gains. This includes highly credible epistemic gains of medium effect size on the overall score as well as axes 1, 3, 3*, 4*.

Comparing the combined *E* scores to the *C* condition, there is a marked difference in the gains made on the EBAPS, including on the overall score and nearly all axes, as shown

FIG. 3. Best estimates of the pre- and post-test scores for the *E* and *C* conditions. Top row: Scores for the overall EBAPS and each of the original axes. Bottom row: Scores for alternative axes identified in Ref. [37]. Error bars indicate 95% HDI for each estimate.

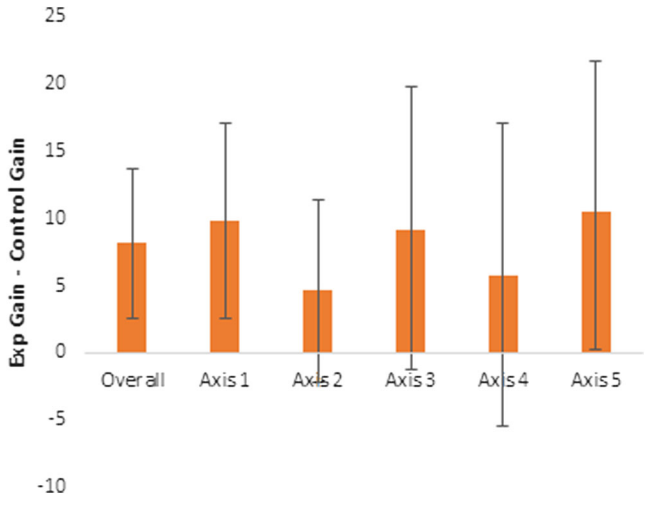

FIG. 4. Best estimates of the difference in score gains made by the E and C conditions. Error bars indicate 95% HDI for each estimate.

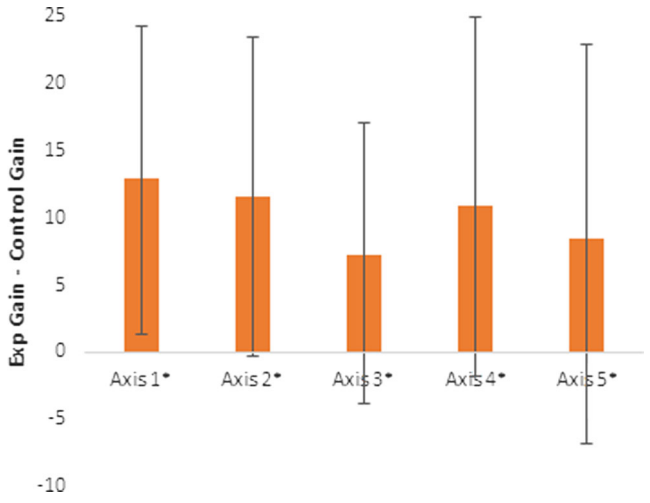

FIG. 5. Best estimates of the difference in score gains made by the E and C conditions, for the alternative axes. Error bars indicate 95% HDI for each estimate.

by Fig. 3 and Table VIII. The difference in gains on the overall EBAPS score is highly credible (>99.9% likelihood) and has an effect size of roughly 0.864, which may be considered large [42]. This is the principal evidence for pronounced epistemic growth by the E condition relative to the C condition. Axes 1, 3, 5, 1*, and 2* also show credible relative gains with medium to large effect sizes, and a few other axes (2, 3*, and 4*) show suggestive relative gains of similar strength. Taken together, the differences in gains between the E and C conditions indicate the E condition made meaningful, credible, and broad epistemic gains relative to the C condition. We next turn to a fuller consideration of these positive effects, their implications for the efficacy of the Bayesian updating activities, and the reasons they may have been produced.

TABLE VIII. Best estimates of the difference in score gains between the E and C conditions, the likelihood that the E condition outperformed the C condition, and the effect size. Superscripts and subscripts denote the associated 95% HDI for each estimate.

Score	Difference of gains	Likelihood $E_{\text{gain}} > C_{\text{gain}}$	Effect size
Overall	$8.2^{13.6}_{2.6}$	>99.9%	$0.864^{1.47}_{0.279}$
Axis 1	$9.8^{17.1}_{2.6}$	99.6%	$0.795^{1.38}_{0.185}$
Axis 2	$4.7^{11.3}_{-2.2}$	91.0%	$0.399^{0.986}_{-0.169}$
Axis 3	$9.2^{19.8}_{-1.3}$	95.7%	$0.511^{1.10}_{-0.070}$
Axis 4	$5.8^{17.0}_{-5.4}$	84.7%	$0.298^{0.887}_{-0.267}$
Axis 5	$10.5^{21.6}_{0.2}$	97.2%	$0.575^{1.18}_{0.002}$
Axis 1*	$12.9^{24.2}_{1.4}$	98.7%	$0.693^{1.33}_{0.084}$
Axis 2*	$11.6^{23.4}_{-0.3}$	97.3%	$0.569^{1.17}_{-0.004}$
Axis 3*	$7.3^{17.1}_{-3.8}$	91.3%	$0.443^{1.07}_{-0.200}$
Axis 4*	$10.8^{24.8}_{-1.8}$	94.5%	$0.474^{1.07}_{-0.090}$
Axis 5*	$8.4^{22.9}_{-6.8}$	87.3%	$0.338^{0.960}_{-0.226}$

V. DISCUSSION

The first consideration must be of the uniformity of the pretest scores between the E and C conditions, as a bias there may threaten the inferential power of the results regarding the efficacy of the Bayesian updating activities. We previously noted that there are significant differences between conditions on the overall pretest score as well as axes 2, 3, and 4*, with the C condition being lower on each. However, as we argue here, we believe that these differences in pretest scores probably do not contribute much, if at all, to the difference in gains observed between the E and C conditions. For example, looking at the three individual sections (C, $E1$, and $E2$) one may note that the overall pretest score estimates are 64.5, 67.5, and 70.2, respectively, which makes for a roughly even spacing of nearly 3 points of difference from one to the next. Yet the post-test scores show increases of similar strength for the two E sections (5.0 and 4.6, respectively), and these gains stand quite apart from the observed decrease for the C section (−3.4). If the differences in pretest scores were significant predictors of the gains made by each section, one would expect a greater differentiation between the $E1$ and $E2$ gains. Instead, we see the two E sections yielding comparable gains despite their own pre-test differences. A similar pattern is observed for the gains on individual axes (both original and alternative). There remains the possibility of some critical threshold for epistemological development, below which students are more likely to demonstrate losses on EBAPS, and above which they may be more likely to produce positive gains, regardless of the curricular materials. There has not been a suggestion of such a threshold in prior studies though, and it would require some very novel and unexpected causal mechanism. In contrast, there is the natural explanation that the sharp two-

level difference in outcomes, which is aligned to the two-level difference in condition and not to the three-level difference in pretest scores, indicates that the gains are due to the Bayesian updating activities.

The size and credibility of the relative gains made by the *E* condition on the overall EBAPS score are both stronger than what was observed in Ref. [33]. This is perhaps not surprising since Ref. [33] featured a control group that used a nontraditional ISLE-based curriculum, which itself exhibited a slight gain on the EBAPS (of $2.6^{7.8}_{-2.9}$). In this study, the control group curriculum was much more traditional and typical of many introductory engineering physics courses, producing a slight negative gain. Yet, it is noteworthy that despite retaining the traditional design for the majority of the lecture, homework, and laboratory work, the *E* sections—modified only with a relative handful of Bayesian updating activities—are able to generate pronounced epistemic gains. This suggests that the efficacy of these activities may be robust no matter what style of physics education environment they are integrated within, and demonstrates that their benefits can be produced with relatively little modification.

A. Effects on axes

It is interesting to consider why the Bayesian activities may have affected specific axes (and alternative axes). Axes 1, 3, 5, 1*, and 2* show credible differences in gains across the two conditions (at the 95% level). Axis 1 (structure of scientific knowledge) assesses whether students believe scientific knowledge consists of disconnected pieces of knowledge such as facts and formulas, or consists of coherent and structured bodies of ideas and information (organized around principles and models). The indirect evaluation activities engage students in the elicitation of prior knowledge (either direct experiences of the everyday world or knowledge of models developed earlier in the physics course) in order to evaluate newer, more tentative knowledge (such as their solution to an end-of-chapter problem). We conjecture that this engagement helps students connect pieces of knowledge, both informal and formal, concrete and abstract, enabling a more holistic and authentic practice of physics by the students.

For example, when evaluating their solutions to an frictionless inclined plane problem, a student knows that when the inclination is 0° the acceleration of any object should be zero, and when the inclination is 90° the object should be in freefall. By testing their solutions to see whether they agree with this knowledge, students are able to use kinesthetic and qualitative experiences from their lives and extract the relevant components of those experiences in order to critically reflect on a problem solution. Moreover, we suggest that these activities help students value the development and deployment of these connections via indirect evaluation, as it is emphasized to them that outside of the classroom one cannot rely on an "oracle" such as an answer key to change one's confidence in the accuracy or utility of a piece of work such as a problem solution. One can only modify confidence in a hypothesis by acquiring new data (i.e., direct evaluation) or by leveraging prior data (i.e., indirect evaluation), and it is generally quicker, cheaper, and more efficient to use prior data when possible.

Axis 1* (structure of science) has some aspects of axis 1 woven within it, but is a bit broader in that the items it includes also consider student beliefs about science as a collaborative enterprise that seeks consistency. That is, there are some sociological and psychological aspects regarding science as a process which are included in this axis, as opposed to axis 1, which focuses on just the interconnected structure of scientific products. The relative boost to gains on this axis may stem from the way the Bayesian updating activities compel students to reflect on consistency between different pieces of knowledge, such as using limit cases to test a proposed solution by reducing it to a simpler situation that they are already familiar with and have strong expectations about.

Axis 3 (Real-life applicability) assesses student perceptions of the utility of scientific knowledge outside of the classroom. Again, by eliciting and valuing the deployment of students' prior knowledge from their everyday lives, the indirect evaluation activities may be responsible for the gains seen on this axis. The direct evaluation activities may also benefit scores on this axis by compelling deeper reflection by students on the actual purpose of the physics labs, and the way theoretical models enable the prediction of outcomes in concrete scenarios. Without the Bayesian updating component, the lab experiments and reports may carry less metacognitive and epistemic value to the students,

Axis 5 (source of ability to learn) assesses students' beliefs about whether the ability to learn and succeed at science is innate or can be cultivated by continual effort. By engaging students in reflective learning, requiring them to practice learning based on self-reflection that elicits relevant prior knowledge, we believe the Bayesian updating activities may develop within students an increased belief in their ability to learn by strategically structured efforts. The gains seen on axis 2* (innate ability vs hard work) are likely due to similar reasons.

In addition to the above axes, similarly credible (likelihood 94.5%) and moderately strong benefits were seen on axis 4* (nature of knowing and learning). This axis is perhaps related to the sophistication of students' epistemic cognition as pictured by King and Kitchener [43,44], as the items included in this axis ask students to evaluate knowledge put forward by the scientific community and also to evaluate their own knowledge, and evokes consideration of the limits of knowing and the presence and importance of uncertainties. These are topics that naturally arise in the Bayesian updating activities, both direct and indirect evaluation types. It seems likely that the experience the students

TABLE IX. Classification of responses to the items in axis 4 according to whether they are consistent with an absolutist or relativist perspective, or neither (neutral). The responses scored as expertlike by the EBAPS are 6A, 28E, and 29C.

Item	Absolutist	Neutral	Relativist
6	A, B	C	D, E
28	A, B	C	D, E
29	D, E	C	A, B

gain at justifying their characterization of an evaluation as confirmatory, disconfirmatory, or null, and their subsequent justification for their chosen value of the R updating coefficient in Bayes's theorem, aids their understanding of the limits of knowing.

Axis 4 (evolving knowledge) shows the weakest relative response between the two conditions. It is not immediately clear why, as one may expect the Bayesian updating activities to help students acquire a sense of how subjective opinions and biases about hypotheses are constrained and eventually converge toward a consensus by repeated evaluation. We speculate that perhaps the lack of repeated testing and Bayesian updating of a single hypothesis (by considering different thought experiments, or running different physical experiments) caused the E condition to shift too far toward relativism, as they may not have acquired sufficient experience in the intersubjective convergence that is achieved via multiple evaluations. To test this, we repeat the analysis of items from axis 4 that was also done in Ref. [33], with responses to the items on this axis being sorted according to whether they indicate a relativist, absolutist, or neutral perspective. The sorting of responses for this analysis is summarized in Table IX.

Comparing the pre- and post-test response rates for each perspective, the C condition response rates for each perspective showed changes of less than 1%, maintaining rates of roughly 45% absolutist, 22% neutral, 33% relativist. The E condition showed slight decreases in the absolutist (−1.5%) and neutral response rates (−2.4%) with a gain in relativist response rates (4.0%). This is similar to what was observed in Ref. [33], although weaker. This offers some support for our conjecture. As noted in Ref. [33] though, the reflective judgment model pictures individuals as progressing through several levels of growth, and it may also be that the students have progressed from absolutist views (roughly corresponding to the "prereflective" level in the reflective judgment model) to relativist views ("quasireflective" level), and this progress cannot be fully detected by axis 4 as it is intended to detect growth toward expertlike views ("reflective" level).

B. Limitations

As alluded to at points in this report, there are a number of factors that limit the inferential power of this study. One threat to internal validity is the difference in class sizes between the E and C sections, with roughly twice as many students in the C section as in $E1$ or $E2$ sections. It is conceivable that the larger enrollment in the C section may have depressed epistemological performance, while the smaller student-to-instructor ratio for the E sections facilitated epistemological gains. Another threat to internal validity is posed by uncontrolled differences in instructors and instructional styles. While broadly similar, they were certainly not identical. Similarly, student demographics between the sections were broadly similar but did differ between the E and C sections and pose an additional threat.

The external validity of our results also faces several limitations. For one thing, the subjects in this study were predominantly white male engineering majors at a regional university. Changes to any one of those demographic and environmental attributes may affect the efficacy of the Bayesian updating activities. Additionally, the E sections in this study included continual instructor-provided feedback on both direct and indirect evaluation activities. If one were to employ the Bayesian updating activities in larger enrollment courses, with teaching assistants or other faculty responsible for such feedback, there may be a reduction in their efficacy.

It is also possible that the more quantitative nature of a calculus-based course enhanced the effectiveness of the Bayesian updating activities, and that weaker gains would be seen if they were added to a traditional algebra-based course. The depth of the quantitative error analysis done in lab reports for a calculus-based course far exceeds what is done in an algebra-based course. Similarly, the extent and power of the limit-case and special-case analyses that can be performed in a calculus-based course are much greater than in algebra-based physics. Both of these differences may cause the epistemological significance of direct and indirect evaluation activities to be enhanced in a calculus-based course.

Finally, while overall instructional time was the same for all sections in this study, it is likely that the gains produced by Bayesian updating activities can be modulated by changes in the duration of lectures and labs. In particular, the course included in this study had two 110-minute lectures per week, plus a 110-minute lab. Other course formats may make either more or less time available for the modeling, discussion, and implementation of Bayesian updating activities. A related factor that may impact the efficacy of these activities would be the production of online videos as resource materials to introduce, explain, and model their completion.

VI. CONCLUSIONS

The incorporation of Bayesian updating activities, including direct evaluation activities in lab reports and

indirect evaluation activities in lecture, homework, and exams, appears to have produced epistemic gains that are credible and of moderate-to-strong effect size. These gains are broad in the sense that they appear across all component axes of the EBAPS, including the alternative set of axes identified by Ref. [37]. They are also of pronounced magnitude, generally matching the gains on EBAPS made by any other curricular materials and course designs in the literature. What makes these materials stand out, though, is that the gains seen here were made with only a modest investment of class time in the Bayesian updating activities, as they were essentially sprinkled into a fairly traditional course design. This suggests that these activity designs may be a way for instructors who lack the resources or training for various nontraditional course designs to nonetheless positively impact their students' development in ways that would otherwise not be possible. Also, when taken together with the prior results in Ref. [33] where additional gains on the EBAPS were made by implementing Bayesian updating activities within a nontraditional course design that had already produced gains of its own, these results indicate Bayesian updating activities have the potential to enrich student epistemological development across a broad range of introductory physics courses, regardless of their underlying design principles.

VII. FUTURE WORK

Future work will explore the incorporation of Bayesian updating within online courses. Virtual and video-based labs are becoming increasingly popular, particularly with the rise of distance learning programs and online course offerings, but the impacts these labs may have on student epistemological growth is unknown, as well as the potential impact Bayesian updating activities may have in that environment. Similarly, the adaptation and use of Bayesian updating activities as components of lecture and homework assignments will be explored.

[1] L. Lising and A. Elby, The impact of epistemology on learning: A case study from introductory physics, Am. J. Phys. **73**, 372 (2005).

[2] D. B. May and E. Etkina, College physics students' epistemological self-reflection and its relationship to conceptual learning, Am. J. Phys. **70**, 1249 (2002).

[3] B. K. Hofe and P. R. Pintrich, The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning, Rev. Educ. Res. **67**, 88 (1997).

[4] M Schommer, A. Crouse, and N. Rhodes, Epistemological beliefs and mathematical text comprehension: Believing it is simple does not make it so, J. Educ. Psychol. **84**, 435 (1992).

[5] K. S. Kitchner, Cognition, metacognition, and epistemic cognition, Hum. Dev. **26**, 222 (1983).

[6] D. Hammer, Two approaches to learning physics, Phys. Teach. **27**, 664 (1989).

[7] E. F. Redish, J. M. Sau, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

[8] R. W. Bybee and B. Fuchs, Preparing the 21st century workforce: A new reform in science and technology education, J. Res. Sci. Teach. **43**, 349 (2006).

[9] R. Gott, S. Duggan, and P. Johnson, What do practicing applied scientists do and what are the implications for science education?, Res. Sci. Technol. Educ. **17**, 97 (1999).

[10] E. Lottero-Perdue and N. W. Brickhouse, Learning on the job: The acquisition of scientific competence, Sci. Educ. **86**, 756 (2002).

[11] S. Duggan and R. Gott, What sort of science education do we really need?, Int. J. Sci. Educ. **24**, 661 (2002).

[12] National Academy of Engineering, *Educating the Engineer of 2020: Adapting Engineering Education to the New Century* (The National Academies Press, Washington, DC, 2005).

[13] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

[14] M. Sahin, Effects of problem-based learning on university students' epistemological beliefs about physics and physics learning and conceptual understanding of Newtonian mechanics, J. Sci. Educ. Technol. **19**, 266 (2010).

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

[16] A. Madsen, S. B. McKagan, and E. C. Sayre, How physics instruction impacts students' beliefs about learning physics, Phys. Rev. ST Phys. Educ. Res. **11**, 010115 (2015).

[17] B. A. Lindsey, L. Hsu, H. Sadaghiani, J. W. Taylor, and K. Cummings, Positive attitudinal shifts with the Physics by Inquiry curriculum across multiple implementations, Phys. Rev. ST Phys. Educ. Res. **8**, 010102 (2012).

[18] V. K. Otero and K. E. Gray, Attitudinal gains across multiple universities using the Physics and Everyday Thinking curriculum, Phys. Rev. ST Phys. Educ. Res. **4**, 020104 (2008).

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

[20] D. Hestenes, C. Megowan-Romanowicz, S. Osborn Popp, J. Jackson, and R. Culbertson, A graduate program for high school physics and physical science teachers, Am. J. Phys. **79**, 971 (2011).

[21] E. Brewe, L. Kramer, and G. O'Brien, Modeling Instruction: Positive attitudinal shifts in introductory physics measured with CLASS, Phys. Rev. ST Phys. Educ. Res. **5**, 013102 (2009).
[22] E. Etkina and A. Van Heuvelen, Investigative Science Learning Environment, Forum on Education of the American Physical Society, Spring issue, 12–14 (2004).
[23] 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).
[24] A. Elby, Helping physics students learn how to learn, Am. J. Phys., Phys. Educ. Suppl. **69**, S54 (2001).
[25] E. F. Redish and D. Hammer, Reinventing college physics for biologists: Explicating and epistemological curriculum, Am. J. Phys. **77**, 629 (2009).
[26] A. E. Lawson, The generality of hypothetico-deductive reasoning, Am. Biol. Teach. **62**, 482 (2000).
[27] A. R. Warren, Impact of teaching students to use evaluation strategies, Phys. Rev. ST Phys. Educ. Res. **6**, 020103 (2010).
[28] A. R. Warren, Evaluation as a means for learning physics, Ph.D. thesis, Rutgers University, 2006.
[29] P. Godfrey-Smith, *Theory and Reality: An Introduction to the Philosophy of Science* (University of Chicago Press, Chicago, IL, 2009).
[30] M. Dancy and C. Henderson, Pedagogical practices and instructional change of physics faculty, Am. J. Phys. **78**, 1056 (2010).
[31] C. Henderson and M. H. Dancy, Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics, Phys. Rev. Phys. Educ. Res. **3**, 020102 (2007).
[32] R. E. Kass and A. E. Raftery, Bayes factors, J. Stat. Assoc. **90**, 773 (1995).
[33] A. R. Warren, Quantitative critical thinking: Student activities using Bayesian updating, Am. J. Phys. **86**, 368 (2018).
[34] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.16.010101 for the Guidelines for Thought Experiments, Guidelines for Lab Reports, the Lab 1 Activity that introduces students to Bayesian updating activities, the answer key for the Lab 1 Activity, the Lab Manual for the first lab experiment of the semester that is done in Lab 2, and the corresponding example Lab Report provided to students as an ideal lab report.
[35] A. Elby, J. Fredriksen, C. Schwartz, and B. White, Epistemological beliefs assessment for physics science, http://www2.physics.umd.edu/elby/EBAPS/home.htm.
[36] H. D. Young and R. A. Freedman, *University Physics*, 14th ed. (Pearson, Thousand Oaks, CA, 2016).
[37] K. Johnson and S. D. Willoughby, Epistemic belief structures within introductory astronomy, Phys. Rev. Phys. Educ. Res. **14**, 010135 (2018).
[38] A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[39] J. K. Kruschke, Bayesian estimation supersedes the *t* test, J. Exp. Psychol. Gen. **142**, 573 (2013).
[40] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/ (2017).
[41] M. Plummer, JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria* (Technische Universität Wien, Vienna, Austria, 2003).
[42] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).
[43] P. M. King and K. S. Kitchener, *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults* (Jossey-Bass, New York, 1994).
[44] P. M. King and K. S. Kitchener, Reflective judgment: Theory and research on the development of epistemic assumptions through adulthood, Educ. Psychol. **39**, 5 (2004).