Framework for evaluating statistical models in physics education research

John M. Aiken[1], Riccardo De Bin,[2] H. J. Lewandowski,[3,4] and Marcos D. Caballero[1,5]

[1]*Center for Computing in Science Education & Department of Physics,*
University of Oslo, N-0316 Oslo, Norway
[2]*Department of Mathematics, University of Oslo, N-0316 Oslo, Norway*
[3]*Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA*
[4]*JILA, National Institute of Standards and Technology and the University of Colorado,*
Boulder, Colorado 80309, USA
[5]*Department of Physics and Astronomy; Department of Computational Mathematics, Sciences,*
and Engineering; and CREATE for STEM Institute, Michigan State University,
East Lansing, Michigan 48824, USA

Across the field of education research there has been an increased focus on the development, critique, and evaluation of statistical methods and data usage due to recently created, very large datasets and machine learning techniques. In physics education research (PER), this increased focus has recently been shown through the 2019 Physical Review PER Focused Collection examining quantitative methods in PER. Quantitative PER has provided strong arguments for reforming courses by including interactive engagement, demonstrated that students often move away from scientistlike views due to science education, and has injected robust assessment into the physics classroom via concept inventories. The work presented here examines the impact that machine learning may have on physics education research, presents a framework for the entire process including data management, model evaluation, and results communication, and demonstrates the utility of this framework through the analysis of two types of survey data.

I. INTRODUCTION

Influenced by the development of machine learning, the statistics community has had discussions on the importance of a model's predictive capability, in addition, or in contrast, to its explanatory power [1–3]. This argument has been extended to the often used statistical inference models in social science [4]. This conversation has moved into the physics education research (PER) community through a recent critical examination of quantitative methods used in PER [5], which included calls to modernize statistical modeling [6]. The framework presented in this paper builds upon that work to produce a method for evaluating statistical studies in PER that is in line with calls for updating current practices in social science fields, such as computational social science and psychology [7].

Traditionally PER has used linear models such as ordinary least squares regression, mixed effects regressions, and logistic regression (e.g., Refs. [8–10]). Linear models

assume that direct combinations of features represent a direct observation [11]. These models focus on determining unbiased estimates of model coefficients in comparison to predictive performance. In the words of Shmueli *et al.* [3], traditional PER models focus on explanatory power rather than predictive power. Models with poor predictive power, as measured by goodness of fit tests, are sometimes still assumed to be transferable and the results generalizable if model parameters are considered statistically significant (e.g., demonstrating via p values and effect sizes). "Hidden" effects are used to explain why there is low explained variance in many models (e.g., Ref. [12]). Using models in this way is not uncommon in other sciences. However, there are calls within physics and other fields to update model assessment and evaluation to include deeper measures of prediction (e.g., Refs. [7,13]).

Much of the recent work in the statistics and machine learning communities has focused on prediction rather than explanation. In the former, community data collected are put into a model that is evaluated on its predicted output. The internal structure of the model is not assumed to be representative of the system being studied [1]; that is, there is not necessarily a correspondence between model structure and system structure and dynamics. The statistics community still considers the model central and highly

values its interpretability. Model performance is evaluated using an independent test set of data [2]. These data are independent; they are completely separated from the data used for model building [14]. Thus, the models are trained using a "training dataset" and then evaluated using a "testing dataset." Models with poor predictive power are rejected in favor of more predictive models [14]. By evaluating the predictiveness, instead of the statistical significance, a relationship between two variables can emerge beyond a "statistical association" [7].

The framework that is presented in this paper is designed to provide a method for establishing generalizable results from educational data mining studies [15]. Its goal is to aid in exploring research questions such as (i) who will change majors in physics to other majors [16], (ii) which students in physics courses need extra help to succeed [17], and (ii) what patterns exist in student preparation that may impact learning at the university level [18]. Producing generalizable results through use of statistical model evaluation will provide an additional tool to the PER tool box. This tool will allow researchers to compare results across similar university environments.

It is important to note that this framework is not designed to test causal models [15], nor is it designed for evaluating latent variable modeling, such as item response theory. While this framework is a good starting point for evaluating these kinds of models too, causal models require more strict verification that is grounded in the theory being tested and cannot rely on solely statistical model evaluations. While this is true in educational data mining studies as well, in an educational data mining study, theory only suggests what variables could be used in a model [15,19].

In this paper, we review how statistical modeling has been done in PER and then describe the current discussions in the statistics and machine learning communities on model representation and evaluation. We then provide evidence that the current quantitative PER work could be advanced by using robust model assessment in line with other quantitative fields. Then, we offer a framework for the PER community that considers data management, model evaluation, and results communication holistically. This framework is grounded in work presented in the recently published PRPER focused collection [5], as well as in modern views from the statistics and machine learning communities. The framework we present is model agnostic, that is, it will work for traditional linear models, as well as more complex "machine learning" models. Finally, we present two applications ("case studies") of the framework: (i) using the Learning About STEM Student Outcomes (LASSO) dataset [20], and (ii) using the data from the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS). These two case studies demonstrate the need for assessing the predictive ability of models beyond what is typically done in quantitative PER.

II. BACKGROUND

Physics education research (PER) is an interdisciplinary subfield of physics that investigates many aspects of physics education [21]. PER uses a wide variety of studies and types of research, including theory building, empirical investigations, qualitative research, and statistical modeling. Quantitative PER has historically focused on investigating how instruction impacts student outcomes such as conceptual learning [22] and attitudinal shifts [23]. Papers in PER that use statistical modeling typically explore data using logistic or linear regressions. Typically, they do not use Cox style regressions (i.e., time-to-event models) or machine learning models (with few exceptions, e.g., Refs. [16,17,24]), although use of these methods is increasing. Early work in PER also generated its own statistics based on empirical observations such as the Hake normalized gain [22].

A. To explain or predict?

A common goal across all science is the understanding and explanation of how a system works. For example, physics has historically had a close connection between the explanation of how a system works and a prediction of the time evolution of the system. Newton's second law provides an explanation of what causes an object to move, as well as excellent predictions as to where an object will end up after some time t. In education research, we rarely have access to such causal relationships that produce excellent explanations of a system, as well as excellent predictions. However, it is sometimes assumed that a model that produces excellent explanations is also predictive [3]. This can conflate the goals of explanation and prediction, which are different aims. This is seen across both the culture of how models are viewed and the practices and choices of the data used for modeling. An explanatory model might rely on a specifically designed survey to, for example, assess a particular attitude towards experimental physics [23] with data from particular types of courses. Whereas a predictive early warning model of student failure in a physics course [17] would not be useful without current data from ongoing courses that are actively streamed into the model.

Breiman *et al.* [1] identified two separate cultures that use statistical models: one culture is grounded in the theory that the stochastic models that generate the data must be approximated, the other culture is grounded in the theory that the model that generates the data should remain unknown. Breiman *et al.* [1] argues that the data generating process is normally complex and statistical models, such as ordinary least squares regression or multi-level regressions, cannot approximate it. In typical machine-learning reasoning, he supports algorithmic methods, in the form of a black-box, which might produce better and different

understanding of the data. In this context, a black-box model is one where the way the model arrives at the solution is unknown. Taking the concept to the extreme, he goes on to state: "the goal [of a statistical model] is not interpretability, but accurate information" arguing that by sacrificing accuracy for interpretability, a researcher sacrifices understanding between the dependent and independent variables." While the advantages of an interpretable statistical model over a black-box solution are evident, in terms of transparency, portability, and generalizability, Breiman's statements should be taken into consideration when highlighting the importance of a model's predictive ability. We argue that a black box that does not give any further insight into the problem is not terribly useful.

Focusing on prediction can lead to the discovery of novel causal relationships, the generation of new hypothesis, and the refining of existing explanatory models [3]. Much of recent literature across social science and statistics has called for a change in the types of models used to predict social systems, the theory and frameworks that motivates the choice of models, and the evaluation methods used to demonstrate prediction. Hofman *et al.* [4] recommend that "current practices for evaluating predictions must be better standardized," arguing that current methods in social science focus too much on explanation and too little on prediction. Hofman *et al.* [4] recognize that there is a fear that complex models may lose interpretability, but points to innovations in recent literature that overcome a loss of interpretability in complex models (e.g., Ref. [25]). The work in PER has not yet fully engaged with this discussion, but work is beginning to discuss the importance of different considerations in conducting quantitative PER [15].

B. Quantitative analysis in PER

Quantitative methods in PER have evolved over time. There is a robust and growing community of researchers in physics education that are thinking about the entire process of quantitative research from data collection and using more sophisticated models to evaluating those models and presenting results. In this section, we discuss a few salient components of the current state of quantitative PER methods and where improvements might be made.

1. Data collection

The PER community has investigated how data preparation and organization impacts observational and experimental results. Both Ding *et al.* [26] and Springuel *et al.* [27] recognize that education data typically are not gathered as interval or ratio data (as is typical in a traditional physics study). Thus, the encoding of data can directly impact the results reported. Springuel *et al.* [27] identifies several types of data that are common in PER and describes the limitations of using each type of data in research. These types include asymmetric and symmetrical nominal data, ordinal data, interval data, and ratio data. These types are

presented on a scale as to how informative they are and what they are able to do. For example, ordinal data contain a finite number of ordered responses (e.g., a Likert scale). In contrast, ratio data create a scale of data from 0 to 1 that represents the presence or absence of some observation.

Missing data can also be an issue in quantitative PER studies [28]. Most statistical models are not capable of handling missing data directly. Data must be either removed or imputed. Rubin [29] describes three types of missing data: (i) missing completely at random, (ii) missing at random, and (iii) missing not at random. Data missing completely at random is when data are missing and there are no mechanisms that can be demonstrated that explain how the data came to be missing. The other kinds of missing data assume that there is some mechanism that may cause the data to be missing (e.g., low performing students are less likely to complete both a pre- and postconcept inventory). When these data are missing in low amounts, it is reasonable to do row wise deletion of data that is missing [27]. Row wise deletion means that data are removed from the dataset for an entire row if any of the columns in a row have missing data.

Nissen *et al.* [28] recommended the use of multiple imputation [30] to fill in the missing data in cases with larger amounts than discussed by Springuel *et al.* [27]. Josse *et al.* [31] demonstrate that in addition to using various imputation methods (including multiple imputation), adding a covariate column that explicitly includes whether data are imputed or not can increase model performance without biasing data, especially when missing data are related to the target variable.

Finally, communication of data context has also been observed to be an issue in PER. Knaub *et al.* [32] found that approximately half of papers published in Physical Review: PER do not describe basic descriptive information about the students being studied such as their demographics, average grades, or institutional descriptors [33].

2. Statistical modeling

Recent efforts have been made to modernize statistical modeling in PER both in methods and in theory. For example, Aiken *et al.* [16] recommend using out-of-sample data to evaluate classification models. Out-of-sample data are data that are randomly selected from the dataset under study that is not used to train the model. Once the model is trained, it is then used to assess model performance (e.g., using mean squared error). Van Dusen and Nissen [6] recognize that much of the data in social science and in PER can have a hierarchical structure that is not picked up by typical linear regressions with only fixed effects. They recommend using multilevel models to calculate the contributions of the hierarchical structure that include random effects. Theobald *et al.* [34] make similar recommendations extending the types of regressions that should be considered in PER (e.g., proportional odds, multinomial, etc.) and
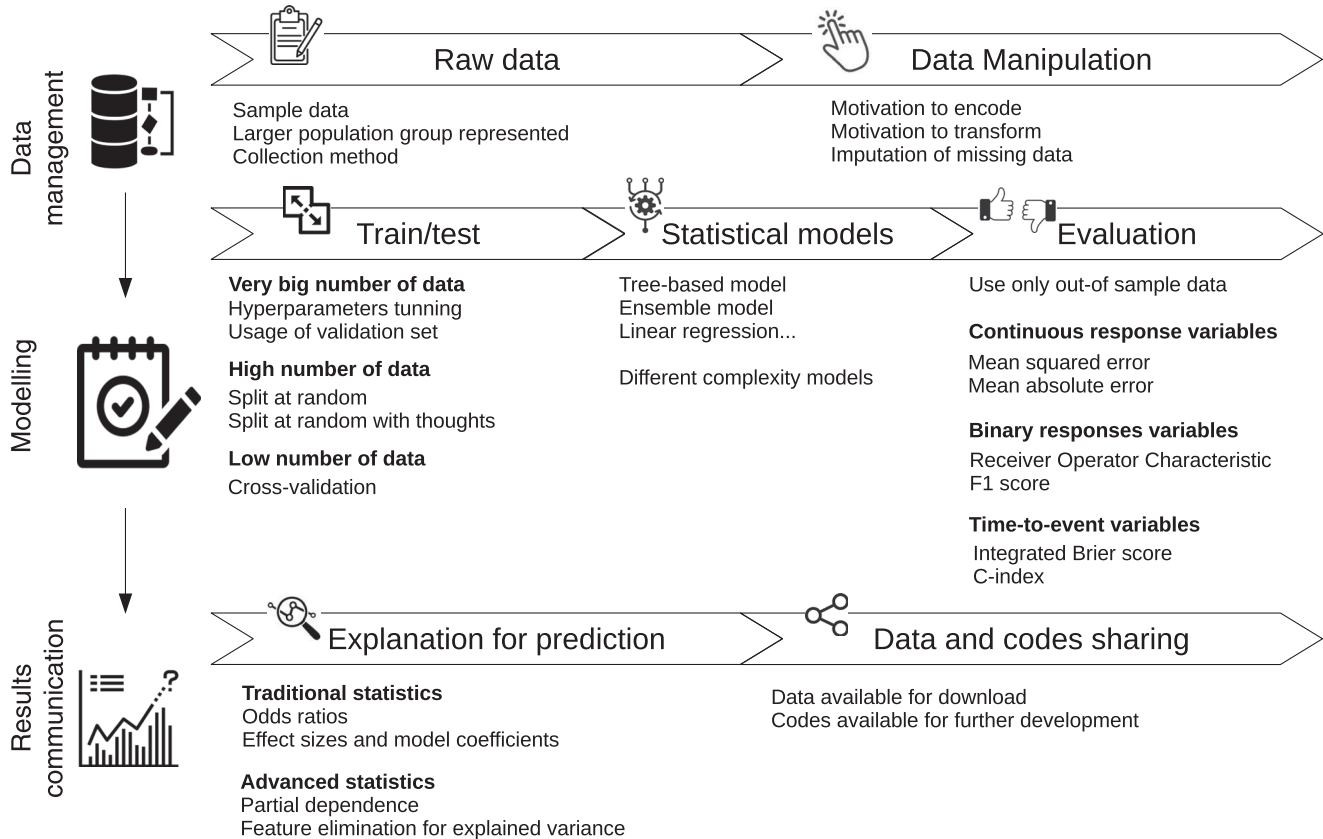
FIG. 1.   A framework for building and evaluating statistical models in physics education research.

provide a toolkit to determine the most effective generalized linear model that a researcher should use to analyze their data. This is in line with much of the work of Bryk and Raudenbush in the 1980s culminating in their textbook [11]. Bryk and Raudenbush [35] identified that "the real need [in education research] is for statistical models that provide explicit representation of the multiple organizational levels typically encountered in educational research." Bryk and Raudenbush [35] identified that learning happens across organizational levels such as an individual student, the classroom, the institution, and even the different times a measurement is taken.

C. The role of theory and domain expertise in statistical model evaluation

The goal of this paper is to present a framework for evaluating statistical models in PER to allow for better generalizability of results. However, we are not suggesting that the only reliable model evaluation comes from whether the statistical model fits the data input into the model or not. It is also important that the data input decisions, and the results that are interpreted, be done in the context of educational theory. Integrating theory and domain knowledge is currently an open topic in the fields of statistics and machine learning. Within PER, domain knowledge and

theory can include, as an example, theories of equity and inclusion [36], theories of learning, such as the role of interaction [22], impact of social interaction and social network [37], neurological and cognitive theories [38], and theories of community integration, such as the communities of practice theory [39]. Statistical modeling allows for a process of rigorous testing of these theories. In reverse, theory allows for strong explanations about when a model fails (e.g., a lack of social network data can impact the overall predictability when examining university student drop out [40]).

III. A FRAMEWORK FOR EVALUATING STATISTICAL MODELS IN PER

The framework presented in this section is aligned with current calls in social science to modernize statistical model evaluation, so that predictive models presented are more robust. Figure 1 is a graphical representation of this framework. Each step in Fig. 1 is discussed below. The framework is shown as a flow diagram to represent each integral step in building a statistical model. This flow diagram represents explicitly each step in the statistical modeling process. It is not meant to represent an ordered process of steps that are performed in succession until completion. In some cases, it is useful to iterate this process

to produce the best models. However, care should be taken to not iterate until results are found (i.e., "*p* hacking" [41]). However, this does not mean that successive tests of different theories and statistical models cannot inform the statistical modeling process and ultimately the results of a research study. Additionally, this process can be transparent through sharing of analysis in an Appendix B, code repositories, and other Supplemental Material [42].

Prior to the collection of any data, researchers must identify their theoretical constraints and research questions. These dictate what data are to be collected at all. For example, if researchers are interested in student learning they might consider a cognitive model of learning and thus they would collect data that would indicate that (e.g., aptitude scores, scores on other cognitive measures, prior course work, amount of sleep, etc.). That is, measurements that all effect cognition. Instead if researchers are interested in identity formation and if someone adopts an identity or not, they could collect other data such as time spent on activities related to that identity, a student's rating of their own competency, etc. Both are welcome in the analytical framework presented in this paper and underlie all of the data and analysis.

A. Raw data

Raw data (see Fig. 1) collected in educational studies should be described in adequate detail. This includes discussing the sample data that are analyzed in the paper and the larger population group that the sample is drawn from. It also means discussing how the data were collected (e.g., in-class, online, for credit, optional, etc.). We recommend following the questions presented in Knaub *et al.* [32] such as "what information regarding the sample is useful for the audience," as well as emphasizing explicit sample descriptions and limitations [33].

B. Data encoding, transformations, missing data

Data collected for any research study is dictated by the research study design and ultimately the research questions being asked. How these data are encoded can impact the results of a model. For example, a student variable could be "passed course" or it could be the student's grade. Encoding and transformation procedures applied to data should be described in adequate detail (see Fig. 1). We do not recommend to state explicitly what each data type is for various kinds of datasets. Instead, we recommend authors motivate the methods used to transform or encode each variable in a model from the type of underlying data. For example, if data are to be transformed using summary data, such as with a z score [43], then the summary statistics (e.g., mean and standard deviation for a z score) should be calculated only for data labeled as training data. These summary statistics can then be used when calculating the z score for the test statistic. Otherwise, data leakage could occur, inflating the model's prediction ability [7]. Data leakage is what happens when data or information from the test set are included in the training dataset. When this occurs, the statistical model can overfit on patterns in the test dataset and present artificially higher predictive results. Additionally, in this step, missing data should be addressed. If applicable, we recommend using an imputation method of the authors' choice (e.g., multiple imputation [44] or K-nearest neighbors [45]) and including a missingness indicator for each column with missing data that has been imputed [31]. In some cases, there may be too much missing data and therefore imputing the data may bias the results [46]. There is no universal solution to missing data. For this reason as well, it is extremely important to report the treatment of missing data in the analysis.

C. Train and test split

This process separates data randomly into training and testing data (see Fig. 1). The training data are used to "teach" the statistical model the patterns in the data. The testing data are used to measure how well the model performs, and should be completely independent from the training data to avoid overestimation of the performance. It is important in this step to acknowledge that some data might be stratified in ways that are inappropriate to completely separate data at random (e.g., randomly picking data across time windows in a study that is attempting to predict time to some event). For example, Aiken *et al.* [16] present a statistical model that predicts if a student enrolled in the physics department will graduate with a physics degree or switch to an engineering degree. In that study, students' high school GPAs had risen over time. One way to adjust model training could be to randomly split data per year instead of randomly across all data regardless of the enrollment year.

When the number of observations is too low to justify a split into a training set and a testing set, more elaborate methodologies can be implemented, for example, cross validation [47]. Cross validation splits the data into several ('K') subsets of approximately the same size: each subset is in turn used as a test set, in which the performances of multiple models trained on the remaining K-1 subsets. The K performance measures obtained are then averaged to obtain a final evaluation of the model. This process allows for the stability of the model to be tested and permit one to use each data point both in the training and test processes.

In addition to data splitting methods for model evaluation, many modern models have hyperparameters that need to be "tuned" to produce a best fit model. Hyperparameters are parameters that describe the structure of the model as opposed to the structure of the data. This could include the depth of a decision tree or the value of a regularization term. When the model fitting procedure requires fixing the value of hyperparameters, such as the number of iterations and the step size for boosting (see Sec. IV A 3 for more details), these must be tuned by using

a third set, called a validation set. In this case, data are split into three groups: training, validation, and testing. The models are trained using the training set for each value of the hyperparameters, the best values are selected using the validation set, and finally the performance of the model is evaluated on the test set. Even if there is no need to tune an hyperparameter, this threefold split can be used to improve the quality of the analysis (model selection and model fitting are performed on independent observations), and therefore a stronger claim of generalizability. Using a validation dataset typically requires having a very large dataset. The split training and validation is mimicked by applying cross validation on the larger set. Cross validation is the same train and test separation process, but abstracted to K separate datasets known as "folds." Then each fold is compared to the other folds to produce average estimates. When cross validation is used instead of the training and test set, a further cross-validation procedure is implemented within the K-1 folds used in turn as the training set.

D. Statistical models

Here, we emphasize that models is plural. The best statistical model might be tree based (e.g., random forests [48] or gradient boosted trees [49]), or it might be an ensemble of models [2], or it might even be a model more familiar to PER researchers, such as a linear regression. The goal is to test models of varying complexity and pick a model that has the highest predictive ability with the lowest complexity both in terms of number of features used and how the model itself works. This is already being done in PER in some studies, typically by comparing models with few features to models with many features and reporting in-sample goodness-of-fit tests. For example, Salehi *et al.* [18] compared different models that predict final exam scores to assess the impact of preparation; Aycock *et al.* [50] compared models that predict a student's sense of belonging in physics due to different experiences. Additionally, more sophisticated models can increase our understanding even when we have lower amounts of data on student subpopulations. For example, using a dataset that had a small racial and ethnic minority population (15%) Aiken *et al.* [40] presented gradient boosted logistic regression models that performed much better than traditionally solved maximum likelihood models when predicting undergraduate student time to graduation. Statistical model choice can impact not only the overall predictability of a model, but also our understanding of a system even when our data on all populations are not equal. It is important to note that the choice of the models to investigate must be done before the data analysis step, but rather in the design of the experiment. Trying all possible methods until one "fits" the data are a way of overfitting (looking for significance) and should be avoided.

E. Evaluate model performance

A model's predictive ability should be evaluated by its ability to predict new data that the model has not seen before [2,7,14]. Using out-of-sample data provides protection against overfitting and places the responsibility of model reliability in its ability to represent the natural world, as opposed to statistical significance established from in-sample data. Standard quantities may be used to evaluate model quality, such as mean squared error or mean absolute error in the case of continuous response variables; area under the receiver operator characteristic curve and the F1 score for binary responses (e.g., for logistic regression); or the Integrated Brier score and C-index for time-to-event data (survival analysis).

A good model fits the training data well, does not diverge much in fit when applied to test data, and has a low complexity. High complexity models typically have a large number of features and are algorithmically complex [2]. They may also suffer from overfitting, with the characteristic situation of small training error and large test error.

F. Explanation

Once a final model has been determined (see Fig. 1), the model can then be used to provide explanations for the prediction. This can include traditional statistics such as odds ratios and effect sizes that are based on in-sample calculated model coefficients. Explanation can also use test data to calculate partial dependence [49] or determine explained variance using recursive feature elimination (e.g., Refs. [16,24]).

G. Using the framework

Ultimately, this framework is only a guide that can be applied and adjusted for each project. For example, a longitudinal study with multiple measurements might require substantial changes in how data are managed and how a model is evaluated. However, several constants should be true for any study that includes predictive statistical modeling. The encoding of data, transformation of data, and handling of missing data should be explicitly articulated. Statistical model evaluation should use out-of-sample data to evaluate model performance. Multiple types of statistical models should be compared and evaluated before presenting results from the model. In the following sections, we demonstrate how to use this framework with two case studies.

IV. ILLUSTRATING THE FRAMEWORK

We have applied this framework to the study of two separate, but common studies in PER. The first study uses the LASSO dataset [20]. The LASSO dataset is a multi-institution, multicourse dataset that includes pre- and

postconcept inventory responses along with other variables. Student postscores are predicted using regression models. The second study uses the Colorado Learning about Science Survey for Experimental Physics [51] to examine differences between students in introductory physics laboratory courses and lab courses that are "beyond the first year" (BFY) of college.

In both cases, the data that are presented have a complex structure, which justifies trying complex models. By complex, we mean that the data may contain nonlinearities, hierarchical effects, and other peculiarities that will lead to poor functional approximation using simple methods such as ordinary least squares. For the LASSO study, first, we try the traditional ordinary least squares solution to linear regression. Second, we attempt to fit a hierarchical linear model using the restricted maximum likelihood method [11]. Finally, we use a gradient boosted trees model [52]. For the E-CLASS study, the problem is framed as a binary outcome variable. Instead of predicting the postscore like before, we predict whether the students are in an introductory course or in a BFY course. In this case, we attempt to solve the logistic regression problem in three ways: (i) with a maximum likelihood estimate method and no hierarchical effects, (ii) a maximum likelihood method with hierarchical effects, and (iii) a gradient boosted method without hierarchical effects. We present these two separate studies as demonstrations of the framework for statistical models that attempt to predict continuous outcome variables and binary outcome variables.

In the following section, we briefly discuss how each algorithm approaches the solution. The methods for each study, such as how the LASSO data or E-CLASS data are transformed, are then discussed in their respective sections.

A. Statistical models

In this study, we use three different methods to derive a regression model. The model is used to estimate or predict the value of a response y, denoted as $\hat{y}$, based on some covariates or predictors x. In the LASSO study, $\hat{y}$ is the post-test score. In the E-CLASS study $\hat{y}$ is a binary variable denoting whether the student is in a BFY course or in an introductory physics lab course. In the LASSO study, the first method we use is a traditional ordinary least squares linear regression. The second is a two-level hierarchical model [11]. The third is a gradient boosted trees model using the Xgboost implementation [52]. In the second study (E-CLASS), we instead use logistic regression models since the outcome variable is whether the student is in an introductory course or a BFY course. Below is a short description of how each method approaches the solution. Then, we describe how the models are evaluated.

1. Ordinary least squares regression

Ordinary least squares (OLS) regression is a popular method used across all fields of science having its roots in the 1800s [53]. Ordinary least squares estimates a vector of coefficients for the model

$$y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon \tag{1}$$

by minimizing the residual sum of squares (RSS) [2],

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} [y_i - \hat{f}(x_i)]^2, \tag{2}$$

where y_i and x_i are the value of y and X for the observation i ($i = 1, \ldots, N$), respectively, and $\hat{f}(x) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$. This vector, $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$, can be obtained directly from the unique solution (not derived here) $\hat{\beta} = (X^T X)^{-1} X^T y$, while $\hat{\beta}_0 = (1/N) \sum_{i=1}^{N} y_i$. In this case, ϵ is assumed to be the irreducible error, i.e., the variability that cannot be explained by the model. It is also assumed that this error is normally distributed with a mean 0 and standard deviation σ^2 [$\mathcal{N}(\mu = 0, \sigma^2)$].

The coefficients of an ordinary least squares model represent how a model variable can impact the outcome variable in two ways: negative or positive (the coefficient sign), the predicted shift in the outcome variable based on a one unit increase in the covariate (the coefficient magnitude).

2. Hierarchical linear model

Hierarchical linear regression models (also called mixed effects models) are another popular method that has seen much use in social science [11] and in PER (e.g., Refs. [54,55]). Mixed effects models allow the inclusion of random effects into the classical linear model, usually something related to clusters of observations. For example, the effect of attending a Catholic school versus a public school may have on math learning [11]. These regressions typically have nested equations that represent the different organizational effects described by Ref. [35]. The simplest version of a random effects model is $y = X\beta + Zu + \epsilon$, where X is the input data, β is the fixed effects, Z is input data with random effects, u is a random effect ($E[u] = 0$, $\mathrm{Var}[u] = \sigma$), and ϵ is the random error. These models cannot be solved analytically like ordinary least squares and iterative methods must be used instead. For a PER treatment of these types of models see Van Dusen and Nissen [6] and Burkholder et al. [56].

3. Gradient boosting

The idea of boosting is to estimate the function $f(X)$ that relates the covariates to the response in an iterative way. At each iteration, called the boosting step, the model is improved by fitting a base estimator $h(X)$ to the negative gradient of a loss function (e.g., the squared loss). The loss

function is a function that penalizes the discrepancies between the outcome of a model and the truth. In the simplest case, this means to iteratively fit a model to the residuals of the model computed in the previous step. At each step, the improvement is kept small (a hyperparameter, called boosting step size, guarantees this) to allow a better exploration of the covariates and, more importantly, to avoid overfitting. Examples of boosting algorithms are the model gradient boosting [57] implemented in the R package `mboost` and the eXtreme gradient boosting [52], implemented in `xgboost`. The latter, in particular, uses, as default, statistical trees [2] as a base estimator.

A full description of gradient boosting can be found in Ref. [2]. The specific optimizations that Xgboost enhances gradient boosted trees with are described in Ref. [52].

B. Model evaluation

It is important to evaluate model performance in order to gauge the confidence we should place in models. Models that have poor fit, such as when R^2 is close to zero, may produce statistically significant results that are not repeatable or are simply due to increases in the size of modern datasets [58]. Model performance for the LASSO study is evaluated visually by comparing predicted distributions (Fig. 3) and through the following metrics (Table III): the mean square error (MSE) and the mean absolute error. Model performance for the E-CLASS study is evaluated by comparing receiver operator characteristic (ROC) curves (Fig. 5). All metrics are computed on a test set.

Here, we do not explore information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). These criteria try to evaluate the prediction error by adding a term to the training error that is supposed to estimate the optimism, i.e., how much the training error underestimates the test error [2]. While accurate for simple models, these criteria tend to be hard to compute for more advanced methods, such as boosting, due to the difficulties in estimating the effective degrees of freedom [59]. R^2 is also not applicable here, as it often has different meaning for different models. Think, for example, to a random effect model where the random effect controls some of the variance of such an effect. Since R^2 represents the explained variance of the model, and the random effect model controls some of the variance through the random effect, then the R^2 will not capture the same explained variance behavior. Despite being a reliable measure in the linear regression model, R^2 should not be used for comparing models different in their nature.

Additionally, we examine the out-of-sample residuals for each model (Fig. 4). Residuals are the difference between the true value and the predicted value from a model. While most summary statistics provide average views of how good the model is in predicting the out-of-sample data

(e.g., mean squared error), residuals plots allow for the examination of local behavior across the range of the outcome variable (i.e., the post-test score).

An assumption of the ordinary least squares regression model is that residuals should be normally distributed and not correlated. When the residuals are not normally distributed and/or they are correlated, this indicates the model missed some behavior in the data. That is, there is an effect that is occurring in the data causing, for example, heteroskedasticity that is not captured by the model's functional approximation of the data. While this assumption is relaxed for other types of models, the presence of correlation within residuals can be useful in analyzing model outputs. Thus, visual inspection of the residuals is often important to determine if models miss underlying structure in the data or overfit on underlying structure. In practice, it is common that residuals have some correlation in real world data [2].

The gradient boosted models presented in this paper use decision tree estimators instead of linear models. Decision tree models do not return coefficients that are easily interpretable like in the case of linear models, instead they return the gain in information due to the use of that variable in the tree. For a gradient boosted model that is a collection of many trees, the average contribution of all the instances of the variable is used to describe its "importance" in the model prediction. For an xgboost model, this value is unitless [52]. Gradient boosted models are evaluated in the same way as linear models using out-of-sample fit statistics.

V. CASE STUDY 1: PREDICTING POST-TEST SCORES IN CONCEPT INVENTORY DATA

Using the LASSO dataset [20], our first investigation looks at predictive regression models of postscores on the Force Concept Inventory and the Force and Motion Conceptual Evaluation. The LASSO dataset is a collection of pre- and postconcept inventory responses collected from multiple institutions typically in introductory physics courses. The data in this paper cover three years of data collection from 2016 to 2018. The LASSO dataset is connected with the national learning assistants' program that aims to include undergraduate students in the instruction of courses [60]. Because this program is teaching focused, LASSO data are likely to be biased by the institutions that select to participate. Institutions that are more likely to invest time and money in teaching and learning are more likely to participate in the LASSO program. Institutions that invest in teaching and learning are likely to see larger learning gains than those that do not [22]. However, the LASSO dataset is a large dataset of validated concept inventory responses, thus, it is a good test bed for the methodology presented in this paper.

TABLE I. Self-reported gender, race, and ethnicity of students in the LASSO dataset. Students may select more than one race option on the survey. In this table the race categories are presented as a sum of the reported within category. Thus the total sum of reported races is greater than the total number of students ($N = 6688$) due to students identifying as more than one race.

	Number of students
Gender	
Female	2422
Male	4266
Race or ethnicity	
Hispanic	1194
White	4381
Black	81
Asian	819
American Indian	60
Hawaiian or other Pacific Islander	60
Race (other)	748

A. Concept inventory data methods

1. Data management

The 2016–2018 LASSO dataset has 13 916 responses for 28 institutions (see Table I). We restrict these data to responses for only the Force Concept Inventory [61] and the Force and Motion Conceptual Evaluation [62] (see Figure 2). This focuses our work on introductory mechanics concepts. We then remove all courses with less than 20 students enrolled and all courses with less than 20 responses. This increases the registered number of students per course for the training and testing datasets. We also remove data for any course that has less than 50% students responding to the pretest or the post-test. Additionally, for students with multiple responses, we use only the student's first response. This reduces the dataset to have 6688 students total across 20 institutions and 90 courses.

In this paper, we have restricted the total variables to 30, including the outcome variable, the post-test score. The variables include two data types: (i) data regarding the student, and (ii) data regarding the course the student attended (see Table II). Student data include the pretest score, background information for students, such as gender and race or ethnicity, the number of pre and post-test questions answered, what year the student is in while taking the course (1st, 2nd, or 3rd or greater), and the time the student spends taking the pre- and post-test. The course variables include the type of concept inventory (FMCE or FCI), the number of students enrolled in the course, the number of learning assistants assigned to the course, the student-to-learning assistant ratio, and whether the course is offered at a community college or not.

The LASSO dataset has missing data [28]. Pretest scores, post-test scores, the number of questions answered for pre- and post-tests, and the time it takes to complete a pre- and post-test all may have missing data. To address this issue, we used the multivariate imputation by chained equations (MICE) method [44] as recommended by Ref. [63] for use on the LASSO data. MICE estimates the missing data per variable that has missing data via a regression algorithm using all other columns. This is repeated for each column with missing data until all columns are filled. This is then
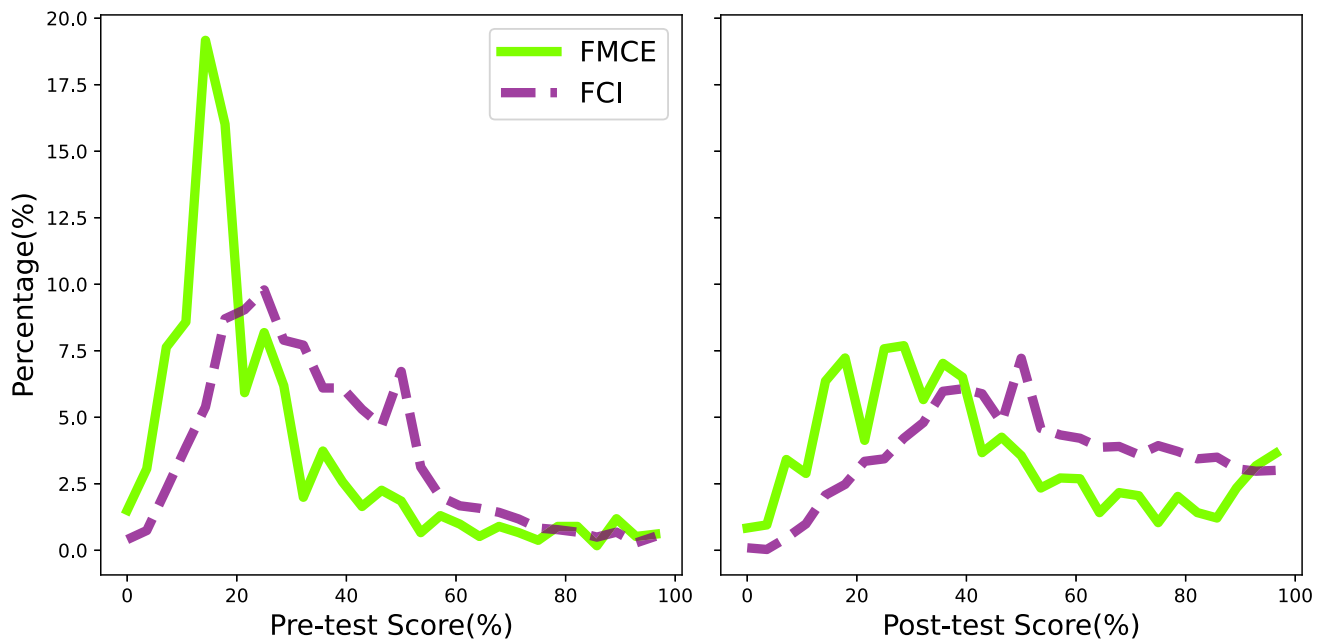

FIG. 2. Distribution of pre- and postscores for FCI and FMCE courses from the LASSO dataset ($N = 6688$). In both the pre- and postscore cases, the distribution of scores are right skewed normally distributed. Data come from multiple institutions across multiple courses ($N_{\text{courses}} = 90$).

TABLE II. Data types at different levels (student or course level). An asterisk (*) indicates that there is also a column in the dataset indicating if the data was imputed or not. No columns for the E-CLASS study were imputed.

Data type	LASSO study	E-CLASS study
Student level	Pre-test Score*	
	Post-test Score*	
	N% Post Answered*	
	N% Pre Answered*	Race or ethnicity
	Log Pre-Duration*	Gender
	Log Post-Duration*	Responses
		to career goals Q's
	Race	
	Hispanic	
	Gender	
	Year Enrolled	
Course level	FCI or FMCE	
	Community College	
	Pre-tests Completed	
	Number of LAs	
	Post-tests Completed	Institution type
	Student to LA Ratio	
	Students Enrolled	
	Course ID	

iteratively fit across a user chosen maximum number of iterations. In this paper, MICE is used to impute the missing data for pretest scores, post-test scores, the number of questions answered for pre- and post-tests, and the time it takes to complete a pre- and post-test. For each variable that includes imputed data, an additional column is created to identify whether the data are imputed or not. This additional column can, in the case when data are not missing completely at random, provide models with information that can aide prediction [31].

2. Model evaluation

In the LASSO study, we use an OLS model, a mixed effects model, and a boosted model. The OLS model uses all of the variables found in Table II. The mixed effects model additionally considers a random effect on the intercept and on the slope. The boosted model uses the xgboost package in python [52] and uses decision trees as estimators. It uses 20 000 decision tree estimators. 20 000 was chosen as it provides a large number of estimators to learn the data without having any one estimator overwhelming the fit. We choose these three models as they each attempt to solve a specific problem and the OLS and mixed effects models are becoming commonly used in PER [6]. OLS models are used to estimate mean effects on outcome variables (in this case, the concept inventory post score). Mixed effects models are used when there is an assumed groupwise variance such that the slopes and/or intercepts per group may be different [6,35]. Gradient
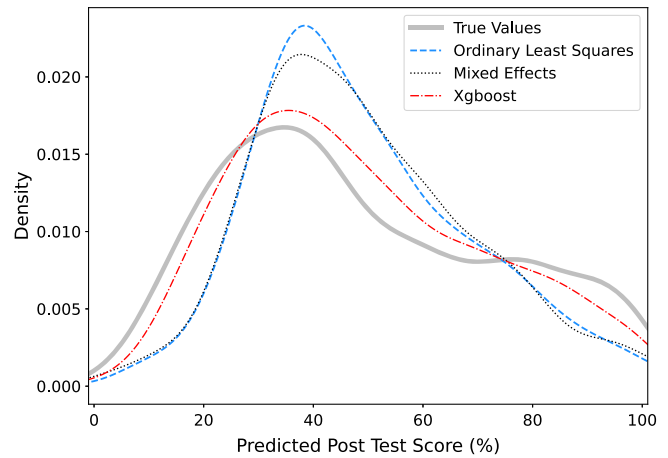

FIG. 3. A comparison of the out-of-sample predicted distributions of post-test scores from the LASSO dataset. Curves represent the kernel density estimates of the predicted post-test scores. The gradient boosted model is closer to the true distribution than the ordinary least squares model and the mixed effects model.

boosting models have been developed more recently [64] and can often better fit data in comparison to maximum likelihood models, such as the mixed effects models presented in this paper [65].

The models are then evaluated using the visual predicted post-score distribution (Fig. 3), the residuals visual (Fig. 4), the mean squared error, and the mean absolute error (Table III). A visual examination of the model fit through predicted post-score distributions and residuals provides a visual representation of the fit statistics (such as mean square error). This visual examination can prevent us from being "fooled" by good fit statistics [66]. The fit statistics provide a summary value that can be compared directly across models. In this case, the lower the mean squared error and the mean absolute error, the better fit the model.

B. Concept inventory data results communication

Using the LASSO dataset, we have trained three separate models: (i) an ordinary least squares model, (ii) a mixed effects model, and (iii) a gradient boosted model. Overall, the gradient boosted model has a lower mean absolute error and lower mean squared error than the linear models (Table III). The mixed effects model has slightly better fit statistics in comparison to the ordinary least squares model.

Visually, the out-of-sample postscore distribution predicted by the gradient boosted model and the mixed effects model is close in shape to the true values (Fig. 3). The ordinary least squares and hierarchical models overpredict scores in the 30%–60% range and underpredict higher scores.

The residuals for all models show some correlation with the true values of the post-test score. The ordinary least
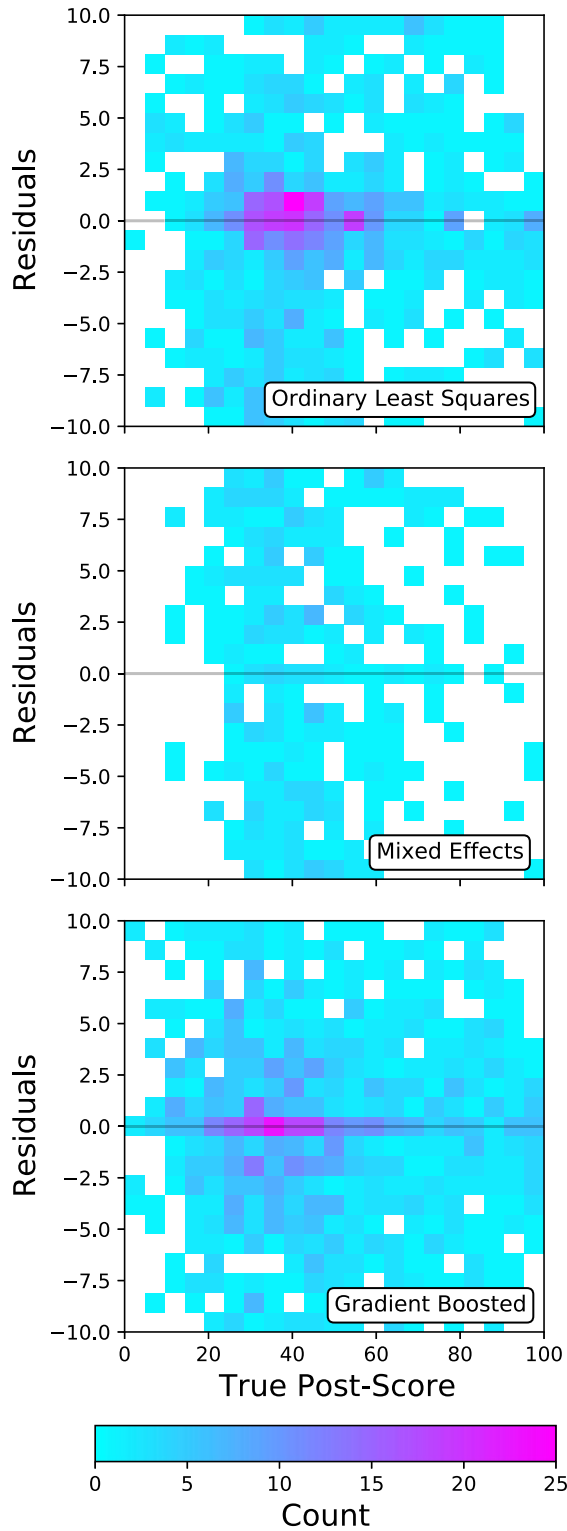
FIG. 4. Residuals plots for the LASSO study. For the OLS model and the xgboost model, the residuals are normally distributed with the majority being within ±10 points of the true value (see Appendix Fig. 6). The mixed effects model does not produce normally distributed residuals.

TABLE III. Out-of-sample model fit statistics for ordinary least squares, mixed effects, and gradient boosted models for the LASSO data study. In all cases, the gradient boosted models have "better" fit statistics in comparison to the linear models. This better fit is unlikely due to the boosted model over fitting the data. This better fit is evidenced by the linear models producing significantly different post score distributions using the out-of-sample data (see Fig. 3).

LASSO Data	Ordinary least squares	Mixed effects	Gradient boosted
Mean absolute error	0.0975	0.111	0.0804
Mean squared error	0.0200	0.0217	0.0152

squares model underpredicts low scores by a wide margin in comparison to the mixed effects model and the gradient boosted model. The ordinary least squares model and the gradient boosted have larger numbers of very small residuals in the range of the most likely scores (20% to 60%). The mixed effects model has a much larger spread of residuals in this region.

VI. CASE STUDY 2: COMPARING STUDENTS IN INTRODUCTORY VS BEYOND FIRST YEAR COURSES

Not all models in PER attempt to predict a continuous dependent variable like we have shown in the previous analysis. In some cases, the dependent variable is binary or even multiclass. For this work, it is common to use logistic regression instead of the traditional regressions. In this study, we use a subset of the data E-CLASS to present an example of a binary classification task in PER and how the framework can be used for such a task. This subset focuses on the future plans questions asked of students in the E-CLASS post survey.

A. E-CLASS methods

1. Data management

The E-CLASS is a survey of student epistemology and expectations surrounding experimental physics. We take a situated theory of epistemological development [67] approach when considering how the survey is used and what research questions can be asked. This theory looks at how students' epistemologies develop in particular social-cultural contexts based on reflections of their experiences. The survey is validated for all levels of university and college students and can be applied across many different lab experiences. To facilitate this process, the E-CLASS survey is administered through a central administration system [23]. As of January 2020, it had been used in 599

TABLE IV. Self-reported gender, race, and ethnicity of students in the E-CLASS dataset separated by level of course (first year and beyond first year).

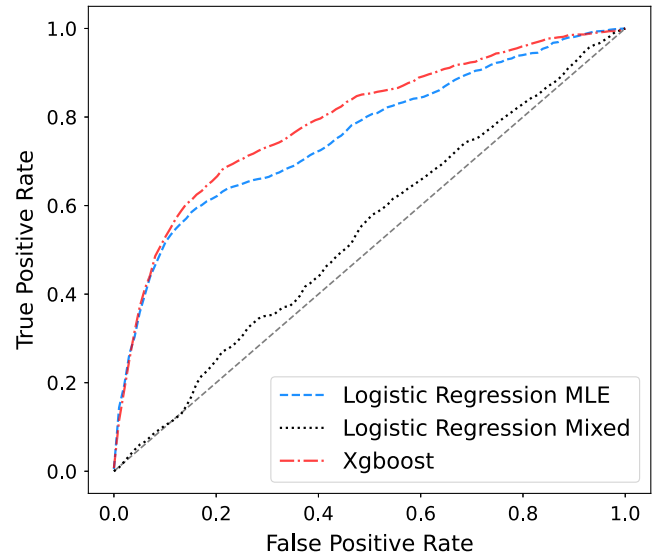|	Intro.	BFY
Gender		
Women	3613	463
Men	5138	829
Other	91	14
Race or ethnicity		
American Indian or Alaska Native	13	2
Asian	1738	217
Black	380	31
Hispanic or Latino	841	48
Multi-race	863	100
Native Hawaiian or other Pacific Islander	31	1
Not reported	542	107
White	4434	800

FIG. 5. Receiver operator characteristic curves for the E-CLASS models predicting whether a student is in an introductory course or a BFY course. The more area under a curve, the better its performance. Overall, the boosted model (AUC = 0.798) performs approximately 2% better than the maximum likelihood model (AUC = 0.773) and considerably better than the mixed effects model (AUC = 0.581). However, the overall scores are still below an excellent discrimination (AUC > 0.8) as identified by Hosmer Jr *et al.* [68]. In this figure, the curves represent the average ROC curve for all 5 cross validation folds.

classes across both introductory lab courses and BFY courses. Data were reduced or removed for several reasons. Students were able to reply to the gender question as a text response instead of picking an option. In this case, these responses were reduced to a third "other" category. In some cases students had replied to the survey more than once, e.g., they took the survey in more than one course. In this case we only used their first response. In some cases there was no description of the type of institution the student attended in the course information survey that instructors filled out. Responses were removed in this instance. Additionally, responses were removed for students who attended 2 year colleges since there are no BFY courses in these institutions. This reduced the overall dataset of matched pre- and postresponses of $N = 19445$ to $N = 10148$ (see Table IV) students. In this study, we focus on the differences between students in introductory and BFY courses.

In this study, we do not use any results from the attitudinal survey itself. Instead, we use demographic information and some of the responses to career plan questions that the students report at the end of the survey. This includes the student's gender, eight questions (see Appendix) concerning the student's preferred career trajectory, the student's race or ethnicity, and the type of institution the student attends (4 year college, master's granting, or Ph.D. granting). We then use these variables to predict whether the students are in introductory or BFY courses.

2. Model evaluation

We apply traditional maximum-likelihood logistic regression, a mixed-effects logistic regression, and a gradient-boosted logistic regression to compare students in introductory courses to those in BFY courses. We compare the predicted output using the receiver operating

curve (ROC) [68]. This curve represents the ratio of true-positive to false-positive predictions of the model given the entire scale of the decision boundary (0 to 1). For example, a threshold of 0.2 would indicate that any student with a predicted probability of < 0.2 would be classified as a student in an introductory course. Students above this line would be classified as BFY students. For the hierarchical model, we add a random term to handle the effect of the institution type.

In this study, we used the advanced form of the "train-test-split" method presented in Sec. III C known as K-fold cross validation [2]. Using 5 folds, we calculate each model and then report the average area under the ROC (AUC) curve (Fig. 5). By using cross validation, we can develop a better understanding of our confidence in the comparisons of the reported AUCs. This is necessary because (a) they end up being low according to acceptable AUC ranges created by Hosmer Jr *et al.* [68], and (b) they are close in value to each other.

B. E-CLASS results communication

Using the E-CLASS dataset, we trained three models to predict whether a student would be in an introductory course or a BFY course when responding to the survey. Using only demographics, responses to the seven-question career-plan question block, and information about the type

of institution, we calculated the ROC curves shown in Fig. 5 from the model classifications (true positive rate and false positive rate). Overall the gradient boosted logistic regression model performs the best (AUC = 0.798), while the mixed effects logistic regression model performed the worst (AUC = 0.581). The fixed effects logistic regression model performs similar (AUC = 0.773), but slightly worse, as compared to the gradient boosted model.

All three models are below an "excellent discrimination" (AUC > 0.8) as identified by Hosmer Jr *et al.* [68]. It is not clear if this can be expected. One must consider the typical demographics of physics lab courses in the US. Introductory lab courses often have students in wide-ranging majors, from physics, to engineering, to those in the life sciences, while BFY courses are typically dominated by physics majors. Students in physical sciences and engineering in the US tend to be overwhelmingly white and male, while those in the life science majors have more women. Additionally, there are issues of systemic sexism and racism in colleges and universities that can decrease the percentage of women and nonwhite students continuing into upper-division courses [69]. It is not clear *a priori* if these general differences in course population are seen in the data we have or can be identified by the models.

VII. DISCUSSION

The framework that is presented in this paper is a culmination of a number of contributions towards statistical modeling in PER. It highlights the need to understand the types of data we collect, how we can represent them, and how we can transform the data [27]. The framework is attentive to the missing data, which is common in social science [28]. The framework allows for the introduction and evaluation of many different kinds of statistical models that can elucidate understanding of educational data [6,16,70]. Finally, the framework aligns statistical model evaluation in PER with calls in social science fields to place a much greater emphasis on evaluating model prediction [7].

A. Data management

Describing datasets used in PER includes how the data are managed. This includes explicit description of how the data are reduced (e.g., dropping data that could be used to identify students) or describing how data are transformed (e.g., how the durations in the LASSO study were transformed via a logarithm). Describing this process allows for interpretation and reproducibility of transforming the raw data into data for a statistical model.

It is common in the social sciences that data are missing and that we need to impute that data in some way [28,46,71,72]. The reason data are missing can be due to both completely random reasons, such as a question simply not being answered on a survey, to purposeful reasons, such

as a subpopulation under study is less likely to participate in a repeated measurement like a pre- and post-test.

In the LASSO data study, we imputed missing data. Examining the feature importances, the gradient boosted model used many of the "is imputed" variables more effectively for prediction than it did the actual data (see Appendix). Approximately half (47%) of the rows in the LASSO dataset had at least one column that was imputed. It could be that having so many data being imputed biases the model much like not imputing data at all can bias results [28]. That is, when so much data are being imputed, the model learns the process the imputation engine follows and that can artificially inflate the model fittedness. Thus, the statistical model then represents what the imputation model did, instead of the nature of the data collected. It may be that there are too many rows with missing data. This is similar to recommendations from biomedical clinical research, which state not to impute data when data are missing at least 40% of the time [46]. It could also be that while the overall LASSO dataset is large, groupwise per course and institution, it has low N values. That is, in both the student groups (e.g., gender, race) and the institutional groups (e.g., single courses) there is a low number of students that represent each subpopulation. Thus, the actual amount of data available to impute student data from the same learning community is low. In this case, it may be wise to examine other methods such as weighting variables instead of imputing variables that have shown to be effective in low-N scenarios [72].

B. Model evaluation

We have presented two research studies in which we demonstrated the utility of examining model fittedness using the statistical model framework presented in this paper. This framework takes an agnostic approach to understanding model success. It is agnostic to whether models are considered direct representations of the data or are simply algorithms that fit data. There is much contention in statistics as to which viewpoint is "better." Bryk and Raudenbush [35] argued for directly representing the organizational levels of an educational environment through random effects in statistical models. This argument forms the basis for the use of most hierarchical mixed effects models. These models have been demonstrated to reject previous results such as the interaction of aptitude and teaching style has on math performance [73]. Ultimately, we must pick the best models that can explain relationships we are interested in, being careful of not searching for the model that explain it by chance. Picking these models must rely on a comparative assessment of many models performance [7]. In many cases, there may be a "multiplicity of good models" [1].

In the LASSO study, we assume there is an institutional random effect due to them being reported in the literature (e.g., Ref. [20]). Therefore, the ordinary least squares

method performs poorly and there is an assumed random effect due to the institution. That is, there is not only a per institution effect that adds some amount to a student's total post-test score, but also this effect has a wide variance that would make the effect variable per student all other variables being equal.

For the E-CLASS study, we used the receiver operator characteristic curve (Fig. 5). The ROC curve represents the ratio of true-positive results to false-positive results. Thus, it helps us estimate model performance in the context of what the model labels as the true value (in our case, being a BFY student). What we see in this study is that if we had relied on an assumed structure of the data as dictated by the mixed effects model, we would arrive at much different results than either the OLS model or the boosted model.

1. Visualizing model results

We have also emphasized in this paper the visual examination of model fits to data, in addition to the fit statistics. It is not a new claim that we must not only examine model fit statistics, but also visually examine the model's predictions themselves. The quartet of Anscombe [66] highlighted the need to examine the data visually. Using a quadrant of four different datasets but with identical fit statistics, Anscombe demonstrated that visual examination of a model's fit is sometimes the only way to determine our trust in the model itself. In this paper, we have demonstrated the same need when statistical modeling is applied within physics education research. In the LASSO case, the gradient-boosted model arrived at a better understanding than the linear models of the dataset. The residuals of the model were examined to better understand the model fit. Using the residuals we can determine how closely the predicted scores are to the actual scores.

2. Boosting

We have presented three models for each study: (i) a traditional linear model, (ii) a mixed effects model, and (3) a boosted model. In each case, the boosted model is better at fitting the data without overfitting. The boosted method likely solves the regression better because the model spends more time learning the peculiarities of each dataset [65]. Gradient boosted models are typically fit using a large number of iterations of adding new learners to the overall base model. The number of iterations is actually an important tuning parameter for boosting, and many properties depend on its correct choice [59].

The models presented here use 20 000 iterations, which produces a post-test distribution closer to the observed (test) data and an overall better fit. No one learner in the boosted model uses all of the variables nor does it use all of the data. By combining the collection of many weak learners, the boosted model is able to capture complex structures of the data, i.e., nonlinear effects, and prevents overfitting (the typical sign of overfitting a model is when

the model predicts training data very well and performs poorly on test data [2]). In particular, the last point is related to the discussion of Sec. VII B on the role of a model: if one wants the results to be generalized outside the specific data set at hand, i.e., have a low prediction error, a boosted model may lead to a much better model than those obtained with traditional approaches.

Ultimately, it is up to the researcher to decide what models they use and why. Researchers who are most concerned with understanding local effects and not producing generalized models are likely to find ordinary least squares an adequate solution for their purposes. However, by restricting the use of which models are available, it might lead to models that require lower confidence in the results or even erroneous characterization of the most important features. Models may also not be able to be generalized beyond a local context. Newer methods, such as boosting, also have the ability to increase understanding of subpopulations that traditionally are poorly fit by statistical models due to lack of data [40]. Ultimately, there is not one "true" model, but many different models that can inform us on the dynamics of students in educational systems in physics.

C. Results communication

It is common in PER to report statistical model results as the coefficients and random effects (if any) of the model. We argue that we can present results in greater detail by also using visualizations of the predicted output.

Models with misestimated variable rankings may lead to departmental policy decisions that are not reflective of the confidence researchers have in the model. Given that 24% of the reviewed papers (Sec. II) presenting statistical models had no examination of model fittedness, this could have a profound effect on results that have been presented in physics education research.

It is often the goal in PER that we are interested in as much the magnitude of an effect as is the direction of that effect. For example, it is helpful to know not only how much a student's pretest score affects their post-test score, but also whether that effect is negative or positive. In this paper, we have not presented these types of results. This is done in order to focus on the examination of prediction to establish model confidence. Modern methods such as boosting can produce these estimates as well depending on the type of learner the boosted model uses (such as a linear learner). These boosted models can also use decision trees (this is what is used in the E-CLASS study). Feature importances of tree-based models do not tell us positive or negative effects only the magnitude of the effect. Instead, tree-based models can utilize methods such as partial dependence [64] or shapely additive values [74] that can provide a more complete picture of each variables effect on the outcome. Both of these model explanation methods explore the entire range of a particular variable and its effect

on an outcome variable such as a post-test score. These measurements are able to explain nonlinear effects (impossible for model coefficients) and are model agnostic. Explaining these methods is beyond the scope of this paper.

D. The role of theory and domain knowledge in statistical modeling

This paper has focused primarily on developing a framework of statistical modeling in PER. The studies presented as examples here are classic studies in PER: survey and concept inventory research. While there is theory motivating these types of studies, the theories were not present in this paper. This is not a tacit recommendation to avoid theory. Theory dictates the research design and the boundary conditions upon which we accept or reject model results. This holds true as well in the reverse. In some cases, theory may fail and the statistics we have used provide an argument for that failure. In either case, it is important to clearly articulate the theory that motivates a study, as much as it is important to articulate a statistical model evaluation process.

E. Recommendations for researchers

Researchers who are building statistical models to answer physics education research questions can build more effective research by following a few recommendations:

1. Describe the dataset, how it was collected, the types of data that are utilized in the study, the context of the data, whether it is from a specific institution or many, and any processes used to impute the data or otherwise alter it [27,32].
2. Predictive modeling studies should use out-of-sample data to evaluate model predictive power prior to relying on model output for explanation [2].
3. Statistics such as the mean squared error and the mean absolute error are appropriate test statistics for comparing model output of prediction models [7]. Examining receiver operator curves (such as shown in the E-CLASS analysis in this paper) are appropriate for classification models [7].
4. Studies should visually examine the predicted distributions and residuals of regression models.
5. Researchers should compare multiple models to produce the best fitted models [1]. These models should include not only different groups of variables, but also different algorithms such as gradient boosting [64] or random forests [48].

While inline with current calls in other fields to standardize model evaluation (e.g., Ref. [7]), these recommendations are not exhaustive. In many cases, there is likely a better fit statistic, visualization, or methodological approach than presented in this paper for the specific research question a physics education researcher is exploring. The recommendations presented here are suggested to be starting points for statistical modeling in PER.

VIII. CONCLUSION

We presented a framework for statistical modeling using predictive models that are presented in physics education research. The framework focuses on explicit articulation of the data management procedures, the model evaluation methods, and the way results are defined and communicated. It presents results that suggest explicit articulation of data handling through transformations, imputations, etc., is important to understand the statistical model outcomes. We highlight the need to assess statistical models using out-of-sample data. Out-of-sample prediction must be used as an assessment of the effectiveness of predictive models. Assessing prediction provides confidence in statistical model results that are otherwise lacking in view of modern statistics [7]. The framework is grounded in recent efforts to critically evaluate statistical methods in PER highlighting this need to assess prediction [5]. We present two complementary studies that justify this need for demonstrating that without assessing prediction, researchers can arrive at erroneous results. In our two example case studies, the boosted models are consistently better than the linear models. While this is consistent with theoretical assessment of boosting from statistics literature (e.g., Ref. [65]), the goal of this paper is not to present boosting to the PER community as a panacea. Instead it is to demonstrate the effectiveness of the presented statistical modeling framework.

APPENDIX A: E-CLASS QUESTIONS

The future plans questions from the E-CLASS post-survey used in this study include the following:

1. Do your future plans include physics graduate school?
2. Do your future plans include a nonacademic science, math, or engineering job?
3. Do your future plans include medical school?
4. Do your future plans include other professional school?
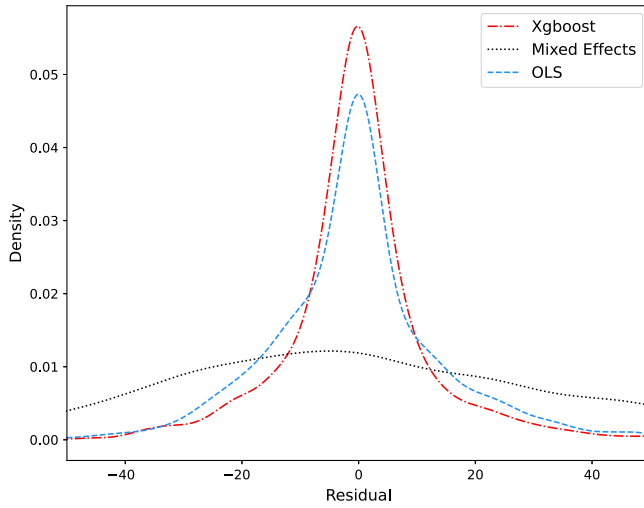5. Do your future plans include teaching K-12 science or math?

FIG. 6.   Residual distributions for the LASSO study.

6. Do your future plans include teaching college science, math, or engineering?
7. Do your future plans include a nonscience, math, or engineering job?

APPENDIX B: MODEL PARAMETERS

1. LASSO study

The LASSO study presented three separate models: (i) an ordinary least squares (OLS) regression (Table V), (ii) a mixed effects regression (Table VI), and (iii) a gradient boosted regression (Table VII). The feature coefficients and importances are presented below. Additionally, a plot of the residual distributions for each model is presented.

TABLE V.   OLS coefficients for the OLS model presented in the LASSO study.

Variable	Coefficient	Standard error	t	$P > |t|$	[0.025	0.975]
Intercept	−33.8622	5.009	−6.761	0.000	−43.681	−24.043
course_id	0.0013	0.001	0.959	0.338	−0.001	0.004
pre_tests_completed	−0.0731	0.011	−6.827	0.000	−0.094	−0.052
post_tests_completed	−0.1110	0.018	−6.309	0.000	−0.145	−0.076
FMCE	−18.6029	2.545	−7.309	0.000	−23.592	−13.613
FCI	−15.2593	2.508	−6.085	0.000	−20.176	−10.343
students_enrolled	0.1507	0.014	11.100	0.000	0.124	0.177
number_of_las	0.1353	0.164	0.825	0.409	−0.186	0.457
student_to_la_ratio	36.7851	7.736	4.755	0.000	21.619	51.952
community_college	0.4356	1.509	0.289	0.773	−2.523	3.394
log_pre_duration	−1.6383	0.797	−2.055	0.040	−3.201	−0.076
log_post_duration	25.8368	0.728	35.481	0.000	24.409	27.264
pre_answered	−0.1613	0.024	−6.837	0.000	−0.208	−0.115
post_answered	0.1560	0.030	5.182	0.000	0.097	0.215
pre_score	0.7692	0.012	62.290	0.000	0.745	0.793
gender	−3.4316	0.445	−7.707	0.000	−4.305	−2.559
hispanic	−4.0267	0.629	−6.405	0.000	−5.259	−2.794
white	5.1159	0.775	6.602	0.000	3.597	6.635
black	−1.1678	1.985	−0.588	0.556	−5.060	2.725
asian	3.7404	0.933	4.009	0.000	1.911	5.569
race_other_reduced	0.6977	0.951	0.734	0.463	−1.166	2.562
year_1st	−2.0976	6.419	−0.327	0.744	−14.682	10.487
year_2nd	−2.9017	6.415	−0.452	0.651	−15.479	9.675
year_3rd_or_greater	−1.8904	6.408	−0.295	0.768	−14.453	10.672
is_imputed_log_post_duration	−0.0805	0.782	−0.103	0.918	−1.613	1.452
is_imputed_log_pre_duration	1.0563	0.824	1.282	0.200	−0.559	2.671
is_imputed_post_answered	−0.9840	0.440	−2.238	0.025	−1.846	−0.122
is_imputed_post_score	−0.5320	0.350	−1.518	0.129	−1.219	0.155
is_imputed_pre_score	−0.9840	0.440	−2.238	0.025	−1.846	−0.122
is_imputed_pre_answered	−0.5320	0.350	−1.518	0.129	−1.219	0.155

TABLE VI.   Mixed effects estimates for the LASSO study.

Variable	Coefficient	Standard error	t value
(Intercept)	−39.1207	13.47173	−2.904
log_pre_duration	−2.0198	0.76716	−2.633
log_post_duration	23.84125	0.71204	33.483
pre_answered	−0.1547	0.02311	−6.695
post_answered	0.18577	0.02882	6.446
pre_score	0.7414	0.01686	43.983
gender	−2.96961	0.44499	−6.673
hispanic	−3.95996	0.61904	−6.397
white	4.6751	0.75039	6.23
black	−2.32269	1.95085	−1.191
asian	2.61889	0.9057	2.892
race_other_reduced	0.16939	0.91586	0.185
year_1st	−4.29582	6.24306	−0.688
year_2nd	−4.86221	6.23259	−0.78
year_3rd_or_greater	−4.5968	6.22859	−0.738
is_imputed_log_post_duration	−1.58494	0.97335	−1.628
is_imputed_log_pre_duration	0.59404	0.8315	0.714
is_imputed_post_answered	−0.62993	0.93741	−0.672
is_imputed_post_score	−0.92841	0.68935	−1.347

Group	Variable	Variance	Standard deviation
course_id	(Intercept)	1.48E + 01	3.84764
	pre_score	6.40E-03	0.080022
students_enrolled	(Intercept)	7.06E-06	0.002658
student_to_la_ratio	(Intercept)	1.53E-03	0.039062
pre_tests_completed	(Intercept)	1.54E-01	0.392324
post_tests_completed	(Intercept)	1.74E-03	0.041762
number_of_las	(Intercept)	2.66E + 00	1.630497
community_college	(Intercept)	4.26E + 00	2.062855
FCI	(Intercept)	1.22E + 02	11.03366

TABLE VII.   Feature importances for gradient boosted model. The feature importance in this table represents the gain [52].

Variable	Feature importance
pre_score	462.408209
is_imputed_post_answered	333.693738
is_imputed_post_score	275.880012
is_imputed_log_post_duration	222.376397
is_imputed_log_pre_duration	153.638948
FMCE	71.928432
log_post_duration	68.916234
black	68.684196
hispanic	59.543799
white	58.978113
post_tests_completed	51.276707
race_other_reduced	40.774619
students_enrolled	39.776663
gender	38.659470
student_to_la_ratio	35.592912
pre_tests_completed	28.820861
number_of_las	27.963471
asian	23.234682
course_id	22.554571
community_college	20.484110
log_pre_duration	16.756483
year_3rd_or_greater	15.908271
year_2nd	14.254940
pre_answered	13.262774
post_answered	12.641225
year_1st	9.258898

2. E-CLASS study

The E-CLASS study presented three separate models: (i) a logistic regression (Table VIII), (ii) a mixed effects logistic regression (Table IX), and (iii) a gradient boosted logistic regression (Table X). The feature coefficients and importances are presented below.

TABLE VIII.   The logistic regression results for the E-CLASS study.

Variable	Coef	Std err	z	$P > \|z\|$	[0.025	0.975]
Gender	0.1637	0.074	2.224	0.026	0.019	0.308
Q53_1	−2.1076	0.085	−24.755	0.000	−2.274	−1.941
Q53_2	0.0992	0.076	1.305	0.192	−0.050	0.248
Q53_3	−0.1475	0.079	−1.870	0.062	−0.302	0.007
Q53_4	0.4417	0.097	4.536	0.000	0.251	0.632
Q53_5	0.0654	0.102	0.644	0.519	−0.134	0.265
Q53_6	−0.1277	0.141	−0.908	0.364	−0.403	0.148
Q53_7	0.2391	0.118	2.031	0.042	0.008	0.470
Q53_8	0.2752	0.115	2.397	0.017	0.050	0.500
Race_Asian	−0.2432	0.100	−2.441	0.015	−0.438	−0.048
Race_Black	−0.6370	0.234	−2.723	0.006	−1.095	−0.179
Race_Hispanic/Latino	−1.0918	0.186	−5.865	0.000	−1.457	−0.727
Race_Other	−0.2945	0.103	−2.860	0.004	−0.496	−0.093
Q15_4 year college	1.0908	0.088	12.414	0.000	0.919	1.263
Q15_Master's granting institution	−0.4627	0.148	−3.120	0.002	−0.753	−0.172

TABLE IX.   Mixed effects results for E-CLASS study.

Variable	Coefficient	Standard error
Intercept	2.3514	0.0366
Gender	0.0379	0.0214
Q53_1	−2.2753	0.0210
Q53_2	0.1316	0.0220
Q53_3	−0.3066	0.0216
Q53_4	0.0661	0.0195
Q53_5	−0.2344	0.0196
Q53_6	−0.3018	0.0189
Q53_7	0.0780	0.0196
Q53_8	−0.0294	0.0192
Race_Asian	0.5905	0.0903
Race_Black	0.2507	0.2267
Race_Hispanic_Latino	−0.0050	0.1702
Race_Other	0.5130	0.0933
Race_White	1.0023	0.0469

Group	Variance	Standard deviation
Institution Type (Intercept)	−0.3489	0.3983
Institution Type (slope)	−1.8655	0.5531

TABLE X.   Feature importances for the gradient boosted model for the E-CLASS study. The feature importance in this table represents the gain [52].

Variable	Feature importance
Q53_1	7.889858
Q53_3	0.735983
Race_White	0.505103
Q53_2	0.437148
Q53_5	0.374478
Q15_4 year college	0.315693
Q53_8	0.089574
Q15_Master's granting institution	0.076935
Q53_6	0.075626
Q53_4	0.071056
Q53_7	0.065396
Race_Hispanic/Latino	0.040842
Race_Asian	0.038410
Race_Other	0.022045
Gender	0.018761
Race_Black	0.005983
Q15_PhD granting institution	0.003979

[1] L. Breiman *et al.*, Statistical modeling: the two cultures (with comments and a rejoinder by the author), Stat. Sci. **16**, 199 (2001).

[2] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical learning: Data mining, inference and prediction* (Springer, New York, 2005), Vol. 27, pp. 83–85.

[3] G. Shmueli *et al.*, To explain or to predict? Stat. Sci. **25**, 289 (2010).

[4] J. M. Hofman, A. Sharma, and D. J. Watts, Prediction and explanation in social systems, Science **355**, 486 (2017).

[5] A. V. Knaub, J. M. Aiken, and M. D. Caballero, Editorial: Focused collection: Quantitative methods in PER: A critical examination, Phys. Rev. Phys. Educ. Res. **15**, 020001 (2019).

[6] B. Van Dusen and J. Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear modeling, Phys. Rev. Phys. Educ. Res. **15**, 020108 (2019).

[7] R. A. Poldrack, G. Huckins, and G. Varoquaux, Establishment of best practices for evidence for prediction, JAMA Psychiatry **77**, 534 (2019).

[8] L. Ding and X. Liu, Getting started with quantitative methods in physics education research, in *Getting Started in PER*, Vol. 2, 3rd. ed. (American Association of Physics Teachers, College Park, MD, 2012).

[9] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? Phys. Rev. ST Phys. Educ. Res. **3**, 010107 (2007).

[10] S.-Y. Lin, J. M. Aiken, D. T. Seaton, S. S. Douglas, E. F. Greco, B. D. Thoms, and M. F. Schatz, Exploring physics students' engagement with online instructional videos in an introductory mechanics course, Phys. Rev. Phys. Educ. Res. **13**, 020138 (2017).

[11] S. W. Raudenbush and A. S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods* (Sage, Thousand Oaks, CA, 2002), Vol. 1.

[12] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[13] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[14] A. Géron, *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media, Inc., Boston, MA, 2017).

[15] L. Ding, Theoretical perspectives of quantitative physics education research, Phys. Rev. Phys. Educ. Res. **15**, 020101 (2019).

[16] J. M. Aiken, R. Henderson, and M. D. Caballero, Modeling student pathways in a physics bachelor's degree program, Phys. Rev. Phys. Educ. Res. **15**, 010128 (2019).

[17] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, Using machine learning to predict physics course outcomes, Phys. Rev. Phys. Educ. Res. **15,** 020120 (2019).

[18] S. Salehi, E. Burkholder, G. P. Lepage, S. Pollock, and C. Wieman, Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics, Phys. Rev. Phys. Educ. Res. **15,** 020114 (2019).

[19] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer, Mining big data in education: Affordances and challenges, Rev. Res. Educ. **44,** 130 (2020).

[20] J. M. Nissen, M. Jariwala, E. W. Close, and B. Van Dusen, Participation and performance on paper- and computer-based low-stakes assessments, Int. J. STEM Educ. **5,** 21 (2018).

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

[22] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66,** 64 (1998).

[23] B. R. Wilcox and H. J. Lewandowski, Students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for experimental physics, Phys. Rev. Phys. Educ. Res. **12,** 010123 (2016).

[24] N. T. Young, G. Allen, J. M. Aiken, R. Henderson, and M. D. Caballero, Identifying features predictive of faculty integrating computation into physics courses, Phys. Rev. Phys. Educ. Res. **15,** 010114 (2019).

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), pp. 1135–1144, https://doi.org/10.1145/2939672.2939778.

[26] L. Ding, X. Liu, and K. Harper, *Getting started in PER-reviews in PER*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2012).

[27] R. P. Springuel, M. C. Wittmann, and J. R. Thompson, Reconsidering the encoding of data in physics education research, Phys. Rev. Phys. Educ. Res. **15,** 020103 (2019).

[28] J. Nissen, R. Donatello, and B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. **15,** 020106 (2019).

[29] D. B. Rubin, Formalizing subjective notions about the effect of nonrespondents in sample surveys, J. Am. Stat. Assoc. **72,** 538 (1977).

[30] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data* (John Wiley & Sons, New York, 2019), Vol. 793.

[31] J. Josse, N. Prost, E. Scornet, and G. Varoquaux, On the consistency of supervised learning with missing values, arXiv:1902.06931.

[32] A. V. Knaub, J. M. Aiken, and L. Ding, Two-phase study examining perspectives and use of quantitative methods in physics education research, Phys. Rev. Phys. Educ. Res. **15,** 020102 (2019).

[33] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16,** 020106 (2020).

[34] E. J. Theobald, M. Aikens, S. Eddy, and H. Jordt, Beyond linear regression: A reference for analyzing common data types in discipline based education research, Phys. Rev. Phys. Educ. Res. **15,** 020110 (2019).

[35] A. S. Bryk and S. W. Raudenbush, Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model, in *Multilevel Analysis of Educational Data* (Elsevier, New York, 1989), pp. 159–204.

[36] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamelá, Toward equity through participation in Modeling Instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6,** 010106 (2010).

[37] R. Dou, E. Brewe, J. P. Zwolak, G. Potvin, E. A. Williams, and L. H. Kramer, Beyond performance metrics: Examining a decrease in students' physics self-efficacy through a social networks lens, Phys. Rev. Phys. Educ. Res. **12,** 020124 (2016).

[38] E. F. Redish, A theoretical framework for physics education research: Modeling student thinking, arXiv:physics/0411149.

[39] E. C. Wenger and W. M. Snyder, Communities of practice: The organizational frontier, Harvard Business Rev. **78,** 139 (2000).

[40] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, Predicting time to graduation at a large enrollment American university, PLoS One **15,** e0242334 (2020).

[41] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, The extent and consequences of P-hacking in science, PLOS Biology **13,** 1 (2015).

[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.020104 for jupyter notebooks that contain codes to perform the analysis in this paper.

[43] J. M. Aiken and M. D. Caballero, Methods for analyzing pathways through a physics major, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 28–31.

[44] S. v. Buuren and K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, J. Stat. Softw. 1 (2010).

[45] N. L. Crookston and A. O. Finley, yaImpute: An R Package for kNN Imputation, J. Stat. Softw. **23,** 16 (2008).

[46] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts, BMC Medical Res. Methodol. **17,** 162 (2017).

[47] J. Shao, Linear model selection by cross-validation, J. Am. Stat. Assoc. **88,** 486 (1993).

[48] L. Breiman, Random forests, Mach. Learn. **45,** 5 (2001).

[49] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 1189 (2001).

[50] L. M. Aycock, Z. Hazari, E. Brewe, K. B. Clancy, T. Hodapp, and R. M. Goertzen, Sexual harassment reported

by undergraduate female physicists, Phys. Rev. Phys. Educ. Res. **15,** 010121 (2019).

[51] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10,** 010120 (2014).

[52] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[53] S. M. Stigler, *The history of statistics: The measurement of uncertainty before 1900* (Harvard University Press, Cambridge, MA, 1986).

[54] J. Bruun and E. Brewe, Talking and learning physics: Predicting future grades from network measures and Force Concept Inventory pretest scores, Phys. Rev. ST Phys. Educ. Res. **9,** 020109 (2013).

[55] M. J. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST Phys. Educ. Res. **10,** 020101 (2014).

[56] E. Burkholder, C. Walsh, and N. G. Holmes, Examination of quantitative methods for analyzing data from concept inventories, Phys. Rev. Phys. Educ. Res. **16,** 010141 (2020).

[57] B. Hofner, A. Mayr, N. Robinzonov, and M. Schmid, Model-based boosting in R: A hands-on tutorial using the R package mboost, Computational Stat. **29,** 3 (2014).

[58] P. E. Meehl, Theory-testing in psychology and physics: A methodological paradox, Philos. Sci. **34,** 103 (1967).

[59] A. Mayr, B. Hofner, and M. Schmid, The importance of knowing when to stop, Methods Inf. Medicine **51,** 178 (2012).

[60] V. Otero, S. Pollock, and N. Finkelstein, A physics department's role in preparing physics teachers: The Colorado learning assistant model, Am. J. Phys. **78,** 1218 (2010).

[61] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30,** 141 (1992).

[62] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66,** 338 (1998).

[63] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14,** 010115 (2018).

[64] J. H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. **38,** 367 (2002).

[65] P. Bhlmann and B. Yu, Boosting with the L2 loss, J. Am. Stat. Assoc. **98,** 324 (2003).

[66] F. J. Anscombe, Graphs in statistical analysis, Am. Statistician **27,** 17 (1973).

[67] W. Sandoval, Science education's need for a theory of epistemological development, Sci. Educ. **98,** 383 (2014).

[68] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression* (John Wiley & Sons, New York, 2013), Vol. 398.

[69] R. E. Scherr and A. D. Robertson, Unveiling privilege to broaden participation, Phys. Teach. **55,** 394 (2017).

[70] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, Using machine learning to predict physics course outcomes, Phys. Rev. Phys. Educ. Res. **15,** 020120 (2019).

[71] C. L. Simons, O. Rivero-Arias, L.-M. Yu, and J. Simon, Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? Quality Life Res. **24,** 805 (2015).

[72] T. Hayes and J. J. McArdle, Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables, Comput. Stat. Data Anal. **115,** 35 (2017).

[73] L. J. Cronbach and N. Webb, Between-class and within-class effects in a reported aptitude, J. Educ. Psychol. **67,** 717 (1975).

[74] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in Advances in Neural Information Processing Systems (2017), https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.