Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictive Modeling for High Impact Active Learning Classrooms

Olive Ross

ogr8@cornell.edu

Cornell University https://orcid.org/0000-0001-6778-6659

Meagan Sundstrom

Drexel University https://orcid.org/0000-0003-4105-215X

N. Holmes

Cornell University https://orcid.org/0000-0001-9141-373X

Article

Keywords:

Posted Date: June 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9955835/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Predictive Modeling for High Impact Active Learning Classrooms

Olive Ross[1*], Meagan Sundstrom[2] and N.G. Holmes[3]

[1*]Department of Astronomy, Cornell University, 410 Thurston Ave, Ithaca, 14853, New York, United States.
[2]Department of Physics, Drexel University, 3141 Chestnut St, Philadelphia, 19104, Pennsylvania, Unites States.
[3]Department of Physics, Cornell University, 410 Thurston Ave, Ithaca, 14853, New York, United States.

*Corresponding author(s). E-mail(s): ogr8@cornell.edu;
Contributing authors: ms5629@drexel.edu; ngholmes@cornell.edu;

Abstract

Over the past several decades, a large body of research has shown that undergraduate science students learn more and more equitably in active learning classrooms; however, the term "active learning" lacks definition and little research has examined which types and combinations of active learning strategies are most effective. In this study, we use a dataset representing over 10,000 students and 24 institutions to create a predictive model that maps classroom time spent on different activities to student conceptual learning. We find that four variables – classroom time spent on lecture, group worksheets, clicker questions, and student questions – are sufficient to reliably predict student learning, as measured by concept inventory scores. We identify one type of class that consistently demonstrates exceptional student learning gains (effect sizes greater than 2): those that spend 10-20% of class time on group worksheets, 20-40% of class time on group clicker questions, and average two or more student questions per hour of class time. We also find that classes which do not utilize group worksheets consistently have learning outcomes comparable to fully lecture classes. These results provide testable recommendations for future controlled studies to investigate effective active learning implementation in undergraduate physics courses.

1 Introduction

Despite lecture being the most common form of instruction for undergraduate science courses [e.g., [13, 25]], an extensive body of work has shown that classes in which instructors lecture for all of classroom time exhibit lower student learning [14], higher student failure rates [14, 20], and less equitable student outcomes across demographic groups [27] than classes in which instructors utilize non-lecture methods. These non-lecture methods are collectively referred to as "active learning" and encompass a wide variety of classroom activities, such as clicker questions, group worksheets, and think-pair-share activities [24, 25]. Accordingly, governmental and institutional bodies recommend that science instructors incorporate active learning into their classrooms [e.g., [19]].

There are, however, myriad definitions of the term active learning, creating ambiguity around how to enact these recommendations [15]. There is also relatively little research pointing to which particular active learning strategies are most effective for student learning and for what portion of class time each strategy should be used to achieve these outcomes. Research studies that directly compare the effects of different active learning approaches on student learning outcomes are necessary to help instructors decide the types of active learning methods to implement and the best ways to implement them [13].

Several existing studies have started to address this goal. One study [26] compared physics student learning outcomes in four research-based pedagogical approaches, finding that the SCALE-UP (Student-Centered Active Learning Environment with Upside-down Pedagogies) pedagogy [4] coincided with higher student learning gains than other commonly used active learning pedagogies. Another study [28] related the extent to which different active learning activities were used during biology classes to student learning outcomes and found that spending more time on group worksheets was positively correlated with student learning. Other research has applied the ICAP (Interactive, Constructive, Active, and Passive) framework [7] to distinguish between different types of engagement, finding that activities promoting higher levels of cognitive engagement are associated with better learning outcomes [8]. While this growing body of work has importantly started to disentangle the relative impacts of different active learning strategies, the impacts of different *combinations* of these strategies has yet to be rigorously analyzed in a way that produces testable predictions about student learning outcomes for future classes.

In this study, we combine data from three extant studies [11, 26, 28]–spanning 24 institutions, 3 scientific disciplines, and class sizes ranging from 11 to 576 students–with new data collected by the authors to identify combinations of instructional strategies that coincide with the largest student learning gains. To do so, we analyze two data sources: classroom activities and concept inventory scores. We characterize classroom activities using the Classroom Observation Protocol for Undergraduate STEM (COPUS) [24]. The COPUS is a structured observation protocol that consists of 25 different codes, or student and instructor behaviors, and observers mark which codes occur in two-minute segments of class time. The COPUS has been used both as an internal assessment of teaching practices [e.g., [3]] and as a research tool [e.g., [10, 25]].

Concept inventories are commonly used by discipline-based education researchers to measure student learning. Concept inventories are, usually, research-validated, multiple choice assessments that measure students' conceptual understanding of core concepts in a given topic (e.g., mechanics or introductory genetics). These assessments are typically administered both at the beginning of a course (pre-test) to determine students' baseline level of knowledge and at the end of a course (post-test) to measure how much students have learned from the received instruction. As in prior work [e.g., [26]], we use effect sizes between pre-test and post-test scores to quantify student learning. We particularly use Cohen's d, a standardized mean difference between the two sets of scores, as this measure can be applied across different concept inventories and student populations [6, 18, 26].

We leverage COPUS observations and student concept inventory scores from 69 undergraduate science courses in this analysis. This dataset represents more than 10,000 students and 40 unique instructors across 24 American and Canadian universities. We use these data to develop quantifiable and testable predictions for which combinations of active learning strategies optimize student learning gains in undergraduate physics courses. Notably, we move beyond the descriptive and/or explanatory analyses conducted in prior studies of the effect of active learning strategies and instead utilize a predictive framework as is common in many other science fields [1, 23]. We create a predictive model that maps class time spent on different active learning strategies to student learning.

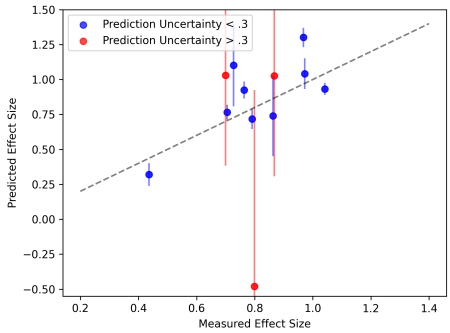
2 Results

We use multiple linear regression to train a model that maps the fraction of two-minute class intervals in which specific COPUS codes were present (approximating the fraction of class time spent on a given activity) to the concept inventory effect size in that class. We apply bootstrap sampling to create a suite of functions rather than a single function, such that the model outputs a distribution of predicted effect sizes rather than a single predicted effect size. The final model prediction is then taken as the mean of this distribution and the prediction uncertainty (due to sample bias) is taken as the standard deviation of the distribution.

This model is limited by the data it was trained on; therefore, we define two criteria that must be met for a prediction to be considered reliable. First, to avoid extrapolation, any predicted effect size outside the range of those effect sizes present in the training data is considered unreliable. Second, any prediction with uncertainty above 0.3 ($\approx$ 10% of the full range of effect sizes present in the training set) is considered unreliable.

We first trained the model on 57 classes from extant data sources [11, 26, 28]. In order to test how accurately the model is able to make predictions for physics and astronomy classes outside of the training data, we collected COPUS and concept inventory data from 12 new physics and astronomy courses. The model was used to make effect size predictions for these classes based on their COPUS profiles and these predictions were compared to the measured effect sizes as shown in Fig. 1. After removing unreliable predictions, the mean absolute error was 0.16, meaning that the

Fig. 1 Effect size predictions on the data collected for the current study from 12 astronomy and physics classes. The model used for prediction was trained on the 57 previously published classes. Classes with prediction uncertainty greater than 0.3 are shown in red.



average difference between measured and predicted effect sizes was ≈ 5% of the full range of effect sizes represented in the data set.

Finally, we trained the model on all 69 classes and used this final model to predict the effect sizes for simulated COPUS data that span the parameter space of the existing data. Predictions that do not meet reliability criteria are hatched in Fig. 2; we leave predictions for classes outside of the range of reliable predictions to future work.

The predicted effect sizes show several novel trends (Fig. 2):

1. Classes that use no group worksheets have consistently small effect sizes, even when other active learning strategies (i.e., group clicker questions) are used (first panel of Fig. 2).
2. In classes spending less than 30% of class time on group worksheets, frequent student questions (≈ 2 or more per hour) are necessary to achieve large effect sizes (first five panels of Fig. 2).
3. Classes that spend 10-20% of class time on group worksheets, 20-40% of class time on group clicker questions, and more than 10% of class time on student questions (≈2 student questions per hour) consistently have large effect sizes (third, fourth, and fifth panels of Fig. 2).

Discussion

Our analysis identifies specific types and combinations of active learning strategies that coincide with large student learning gains in undergraduate science courses. In this section, we discuss our findings in light of previous work and, where relevant, suggest avenues for future work. First, we find that in-class group worksheets are strong predictors of student learning, a result that agrees with the findings of Weir and colleagues [28]. As group worksheets are inherently interactive activities, this result also aligns with the ICAP [7] framework's assertion that constructive and interactive activities lead to the highest learning gains.

Second, our model indicates that student questions are an important driver of student learning, suggesting that instructors should encourage students to ask questions

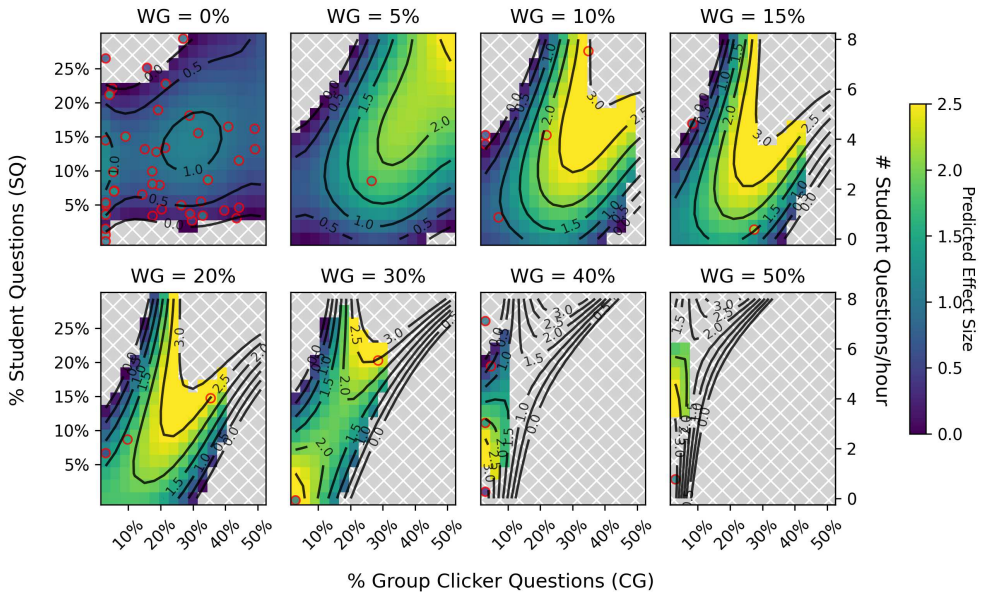


Fig. 2 Heatmaps of predicted effect sizes from simulated COPUS code data by fraction of class time spent on Group Worksheets (WG), Group Clicker Questions (CG), and Student Questions (SQ). The amount of lecture (Lec) is calculated as $Lec = 1 - CG - WG$, assuming that all class time not spent on clicker questions or worksheets is spent lecturing. Student questions (SQ) are almost always co-coded with lecture (Lec) and therefore are not present in this approximation. Because student questions tend not to take up a whole two-minute COPUS interval, the right y-axis translates from percent of class time spent on SQ to average number of student questions per hour of class. Gray hatched regions are predictions that have high prediction uncertainty (bootstrap error >0.3) and/or predicted effect size values outside the range of effect sizes present in the training data. Training data is binned to the nearest WG value and plotted over the heatmaps, shown outlined in red and filled in with colors representing the measured effect sizes.

in front of the whole class. While past work has investigated the impact of student questions on the asker and the instructor [9, 21], little work has been done to investigate the impact of individual, in-class, student questions on the learning outcomes of the whole class. We also find that more class time spent on student questions is correlated with a decrease in the width of the distribution of student scores from pre-test to post-test (see Fig. 3), potentially indicating student questions as a driver of more equitable student learning. The mechanisms of a single student's question impacting the whole class's learning should be explored in future work.

Third, the model predicts one "optimal" combination of active learning strategies that should be tested in controlled studies. Classes that spend 20-40% of class time on group clicker questions, spend 10-20% of class time on group worksheets, and have 2 or more student questions per hour consistently have high learning outcomes, even when all other class time (i.e., up to about 60% of class time) is spent lecturing.

Lastly, the results indicate that less class time spent on lecture is not necessarily better for student learning, as suggested by some prior work [e.g., [13]]. While it is true that classes that are almost entirely lecture-based consistently yield lower

learning outcomes, according to our model, there are types of majority-lecture active learning classrooms that demonstrate high student learning outcomes. However, we find that classes which rely on group clicker questions as the only form of active learning have consistently low effect sizes, comparable to those of fully lecture classes. This combination of results is particularly surprising given the abundance of research studies evaluating the effectiveness of clicker-based pedagogies, such as Peer Instruction, as compared to traditional lecture courses [e.g., [5, 12]]. Future work should seek to disentangle these conflicting claims.

Several limitations of this study also prompt future investigations. An undergraduate science class has many more variables than are captured in our model. First, while there are many potential metrics and outcomes on which to assess the quality of an undergraduate science course (e.g., problem solving skills, effective collaboration, self efficacy, critical thinking), in this study we focus solely on conceptual learning. Future work should examine the differential impact of active learning strategies on other student outcomes. We also do not account for activities that occur outside of regularly scheduled lecture time (such as homework, quizzes, projects, laboratory sections, discussion sections, or time spent studying) or for in-class variables not captured by the four COPUS codes (CG, SQ, Lec, and WG) included in this study. Even within the four variables considered in this work, the sparsity of our data in certain regions of this parameter space severely limits our ability to make reliable predictions. While one of the highest effect sizes represented in our dataset was from a class that spent 100% of class time on group worksheets, we were unable to make reliable predictions for classes that spent more than 50% of class time on group worksheets due to the sparsity of the data in that range (see Fig. 3). The success of that singular class suggests that classes spending a majority of time on group worksheets are another promising type of active learning class that should be statistically evaluated in future work.

Our COPUS data also represent average rates of code occurrence over two to nine observed class sessions per course. While a sample of two to three classes is generally considered a representative sample of classroom activities [16, 25], the particular class sessions observed in our dataset may not represent typical or average classroom behaviors across the entire course. Additionally, because COPUS coding occurs in two-minute intervals, an occurrence of a code of any duration within a given interval is marked as being present for the full interval. This feature of the protocol may lead to an overestimation of class time spent on codes. Future work should consider using continuous coding instead of two-minute intervals to capture more accurate fractions of class time spent on different activities.

Notably, this work focuses on the statistical accuracy of model predictions and not the theoretical underpinnings or causality of the relationships represented by the model. Future work should evaluate whether the presence of these activities directly impacts student learning, or if they serve as proxies for other instructional or interpersonal variables.

Finally, our data are highly varied across class size, student demographics, student academic level and preparedness, and instructor experience. While this diversity enables the broad generalization of this work, there may be hidden variables among these classes that cause the independence of this dataset to break down. There may

be selection bias effects both for instructors who volunteered their classes to participate in these studies (e.g., instructors who are discipline-based education researchers themselves may be more likely to volunteer and may exhibit larger student learning gains than other instructors [2]) and for students who completed the pre-post surveys (e.g., high-performing students are more likely to participate in research studies than low-performing students [17]). Future work should examine these and other factors that may impact the model results, such as by further comparing the model's prediction accuracy across different populations of students, disciplines, course levels, and other contextual and demographic variables.

Methods

Data

There are two sources of data in this study: those sourced from extant studies and those collected for the current study (Table 1). We used extant data from three different published works [11, 26, 28]. We identified these works by first searching for all papers that have cited [24], the paper that introduces the COPUS, and then identifying those that collected both COPUS and concept inventory data. From these works, we then selected those with publicly available data containing the fraction of total class time spent on COPUS codes and concept inventory means and standard deviations. We also contacted the authors of papers without publicly available data; however, none of those authors were able to provide the required data for our analysis. The dataset from existing studies represents 57 class sections, more than 8,000 undergraduate science students, and 24 American and Canadian Universities. To supplement these data, we also collected COPUS and concept inventory data from 12 introductory physics and astronomy class sections containing more than 1,900 undergraduate students at a private, R1 university in the United States.

Quantifying Classroom Activities with the COPUS

The COPUS data were all collected either through live classroom observations or video recordings. To apply the COPUS to a live or video-recorded classroom observation, a trained coder indicates whether or not each COPUS code was present during each two-minute time interval [24]. The COPUS code quantities reported here are the percentage of total two-minute time intervals observed in which the given code was present. Multiple COPUS codes can be applied in each time interval; therefore, the sum of all COPUS code quantities is more than one.

We include four COPUS codes in this study: Lec (instructor lecturing), WG (students working on worksheets in groups), CG (students working on clicker questions in groups), and SQ (student asking a question). We selected these COPUS codes by first identifying all COPUS codes for which data from extant studies were reported: Lec, WG, CG, SQ, PQ (instructor posing questions), OG (other group work), and CQ (instructor posing a clicker question). We removed PQ and OG from our analysis because they are quite broad (i.e., PQ may refer to the instructor asking any question or assigning any kind of groupwork activity and OG may refer to different kinds

of groupwork, such as laboratory experiments or solving problems at whiteboards) and, therefore, are not aligned with the goal of this work to provide testable predictions about specific classroom activities. Additionally, because of the high covariance between CG and CQ, it was necessary to remove one of these variables. We trained the model for both cases and found significantly higher prediction accuracy when using CG; therefore, CG was retained in our analysis and CQ was removed.

Fig. 3 shows the prevalence of values for each COPUS code in the dataset as well as their relationships against the change in concept inventory score means (gain), the change in concept inventory score standard deviations, and effect sizes. Notably, our results do not include predictions for classes with $WG > 50\%$ due to the sparsity of the data in that range. Because of the complex, higher order relationships we find between these variables, it is not surprising that few COPUS variables have notable linear correlations with the selected summary statistics.

Quantifying Student Learning with Concept Inventories

We measure student learning using the effect size Cohen's d [18] between pre-test scores and post-test scores on concept inventories, defined as:

$$d = \frac{m_{post} - m_{pre}}{\sqrt{\sigma_{pre}^2 + \sigma_{post}^2}/2}, \tag{1}$$

where m_{pre} and m_{post} are the mean student scores on the pre-tests and post-tests, respectively, and σ_{pre} and σ_{post} are the standard deviations of the pre-test and post-test scores, respectively. This metric provides a dimensionless measure of student learning in each course, allowing us to combine student scores across a range of disciplines, courses, student populations, and concept inventories. The ability to compare scores across concept inventories is particularly important in this work, with 28 different concept inventories represented across our dataset. Fig. 4 shows the relationships between concept inventory mean pre-test scores, mean post-test scores, and effect sizes present in the data. The nature of this work necessitates the use of effect size as a summary statistic, however, as demonstrated in Fig. 5, effect size is strongly correlated with gain. This may aid in interpretation for readers with more familiarity with gain as a summary statistic. We removed three classes from the dataset because they had pre-test scores over 60% and high pre-test scores limit the range of possible learning gains, creating ceiling effects that lead to non-normal distributions of scores.

Modeling

In this section, we describe our development of a predictive model that maps four COPUS codes – Lec (lecture), WG (group worksheets), CG (clicker question group work), and SQ (student questions) – to effect sizes. Source code is available at https://doi.org/10.5281/zenodo.19598369. Model weights were fit using the ordinary least squares (OLS) module from the STATSMODEL [22] package in Python. Because the aim of this work is predictive by nature [1, 23], we used predictive metrics for model selection, namely the small sample adjusted Aikake Information Criterion (AICc). AIC is a measure of the predictive power of a model that quantifies the amount of

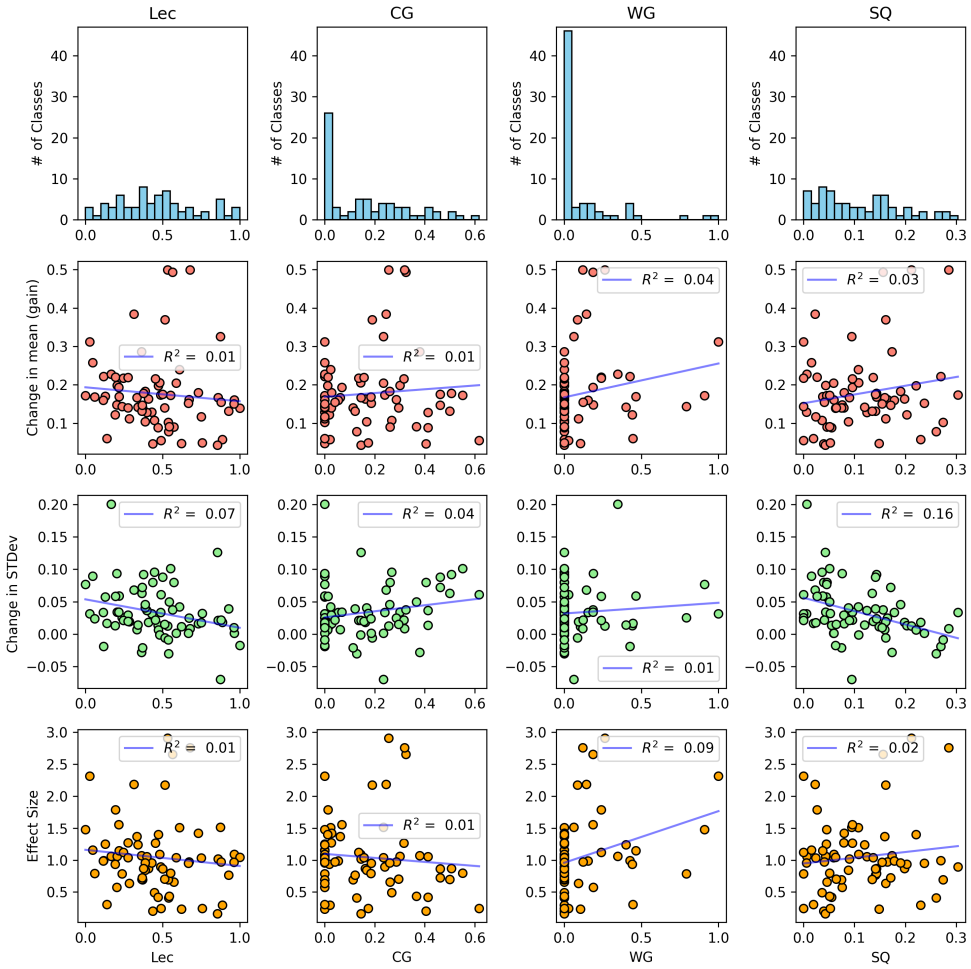


Fig. 3 COPUS Variables vs. Summary Statistics. Distribution of classes with respect to COPUS values (row 1), plots of COPUS code values against the difference between mean post-test scores and mean pretest scores, or gain (row 2), plots of COPUS code values against the difference between the standard deviation of posttest scores and the standard deviation of pretest scores (row 3), and plots of COPUS code values against effect sizes (row 4).

information lost when the model is applied. The measure rewards goodness of fit by maximizing a likelihood function and penalizes model complexity (e.g., large numbers of predictor variables) to avoid overfitting.

The relationship between classroom activities and student learning outcomes is complex. Some activities may have simple linear relationships with learning (e.g., lecture time is negatively correlated with effect size) while others may have more complicated relationships (e.g., effect size peaks at 20-30% of class time spent on group clicker questions). In order to allow for these non-linear effects, we included the fraction of class time spent on the four COPUS codes of interest (Lec, WG, CG, and

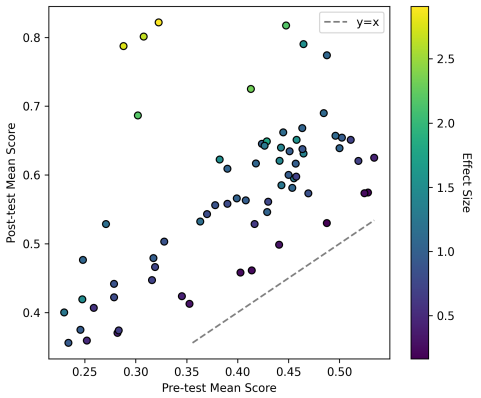


Fig. 4 Mean pre-test scores vs. mean post-test scores colored by effect size. The dashed line marks $y = x$, so a class falling on that line would have zero learning gain. Pre-test standard deviations range from 0.07 to 0.28, while post-test standard deviations range from 0.12 to 0.31.

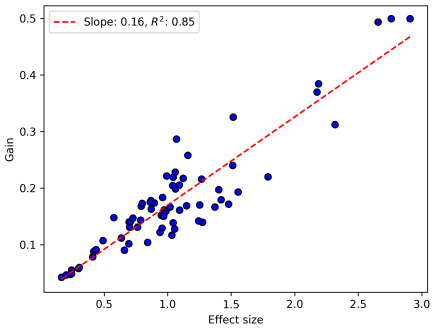


Fig. 5 Relationships between effect sizes and gains (mean post-test scores minus mean pre-tests scores). The dashed red line denotes a best fit line with a slope of 0.16, indicating that an effect size of one is comparable to a gain of 0.16.

SQ), their squares, cubes, and quartics (e.g., Lec^2, Lec^3, Lec^4), and all second and third order cross-terms (e.g., $WG * CG$, $WG^2 * CG$, and $WG * CG * SQ$), as features in the model.

We ran the OLS fitting algorithm on all of these features and computed the AICc. Then, one feature was removed, the fitting algorithm was run again, and the AICc was computed. If the AICc decreased when that feature was removed (meaning the model fit improved), then that feature was removed from the model. This was repeated for all features to select features for the final model. The order in which this feature removal is performed has an impact on the features that are ultimately selected. For this reason, this entire feature selection process was repeated 100,000 times with different feature orders. The feature sets with the 10 lowest associated AICc values were retained. The number of retained features across the 10 feature sets varied from 22 to 24 (out of 39 total).

Quantifying Model Uncertainty

The dataset used to train our model is limited (i.e., it does not include every possible type of class); therefore, the predictions we are able to make are limited. To quantify how sensitive the model predictions are to sample bias, we performed bootstrap sampling. We randomly selected 1,000 different subsets of 85% of the data and trained the model on each of those subsets for each of the 10 retained feature combinations mentioned above. Those weights were averaged over feature sets and the final model includes 1,000 sets of weights. Therefore, the model outputs a distribution of 1,000 different predictions for each set of COPUS codes. The final prediction is taken as the mean of that distribution and the model uncertainty on that prediction is defined as the standard deviation of that distribution. We defined all predictions with uncertainty above 0.3 ($\approx$10% of the full range of effect sizes) as unreliable. We do not make any claims about unreliable predictions other than to identify them as areas of possible interest for future work.

Out-of-sample predictions

The goal of this work is to predict effect sizes for classes outside of the training sample. To confirm that our final model can accurately predict effect sizes for classes outside of the training set, we performed three tests. First, we trained the model on the extant dataset, used that model to make predictions for the newly collected data, and evaluated the model ability to make accurate predictions for new physics and astronomy classes as described in 2.

Second, we used a type of data partitioning called a Leave-One-Out (LOO) test, where all data points (in this case classes) except one are used in training. The model is then used to predict the effect size in the left out class to assess how well the model can predict effect sizes for classes outside of its training set. This process is repeated for each class in the extant data set (57 classes), providing information about how well the model will predict effect sizes for new classes (Fig. 6). When we remove predictions that are outside the range of reliability (i.e., those with uncertainty above 0.3), the mean absolute error is 0.18. This means that, on average, the model makes predictions within 0.18 of the measured effect size in a given class when that class is not included in the model's training data.

Lastly, we evaluated the robustness of the model across class-level variables by distinguishing between the fields and class sizes for the courses in the extant data set during the LOO tests (Fig. 7). When we remove predictions outside the range of reliability, the mean absolute errors on predictions are 0.16, 0.19, and 0.13 for physics, biology, and astronomy respectively. Similarly for class sizes, when unreliable predictions are removed, we find mean absolute errors of 0.16, 0.19, and 0.18 for class sizes less than 50 students, class sizes between 50 and 150, and class sizes over 150, respectively. Note that the values used here for class size are the number of students who both completed concept inventories and consented to participate in the studies, therefore these values are systematic underestimates of the true total class sizes. We do not find any significant systematic effects for model predictions as a function of field or class size.

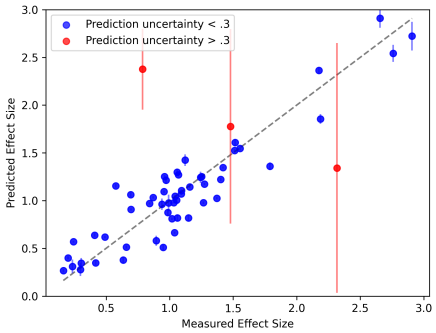


Fig. 6 Results of Leave-One-Out tests for 57 previously published classes. Each class is removed from the dataset, the model is trained, and a prediction is made for the left-out class. Classes with prediction uncertainty higher than 0.3 are shown in red.

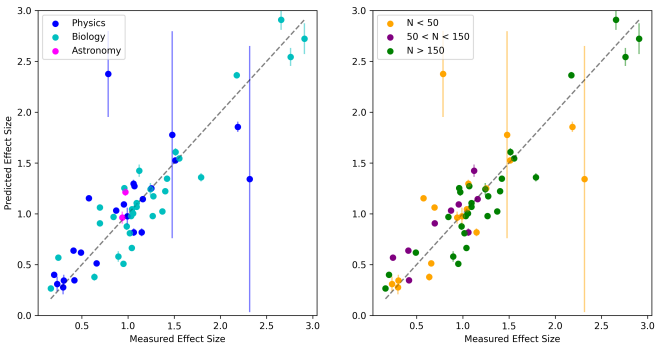


Fig. 7 Results of Leave-One-Out tests for 57 previously published classes by field (left) and class size (right). Each class is removed from the dataset, the model is trained, and a prediction is made for the left-out class.

Final Model Training and Generation of Results

After ensuring that the model can reliably predict effect sizes for data outside of its training set, the model was ultimately trained on data from all 69 classes (i.e., both extant data and new data collected for the current study). This model was used to create the predictions described in the main text (Fig. 2). Due to the 4-dimensional nature of the COPUS data and the associated limitations of plot representation, we chose to exclude the lecture variable from these plots. Instead, at each point on each heatmap, we calculate the lecture time as $Lec = 1 - WG - CG$, assuming that all class time not spent on group worksheets or clicker questions is spent on lecture. We excluded student questions from this equation because they are almost exclusively co-coded with lecture. This estimate functions as an upper bound on the amount lecture time in a given class as demonstrated in Fig. 8, but notably leaves out the variation

in time spent on classroom activities other than lecture, group worksheets, and group clicker questions.

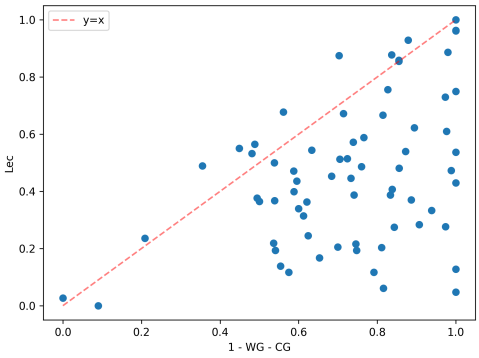


Fig. 8 *Lec* vs. $1-WG-CG$ from the training data. This relationship demonstrates that the equation we have chosen to estimate the amount of lecture in a classroom acts as a plausible upper bound on lecture time.

Acknowledgments

We are grateful to the instructors who allowed data collection in their courses and the researchers who shared their data publicly. We thank the members of the first author's doctoral committee as well as Tim Steltzer, Peter Lepage, Eric Brewe, and our collaborators at Cornell University and Drexel University for their guidance and feedback at various points throughout this project.

Funding:

O.R. discloses support for the research of this work from the U.S. National Science Foundation Graduate Research Fellowship (DGE – 2139899). N.G.H. and M.S. declare no relevant funding.

3 Author contributions:

O.R. conceptualized the study. O.R. and M.S. curated the data. O.R. performed the formal analysis and investigations. O.R., M.S., and N.G.H. defined the methodology. N.G.H. and O.R. managed project administration. O.R. designed the software. N.G.H. supervised the project. O.R. conducted validation and performed visualization. O.R. wrote the original draft, which was reviewed and edited by M.S. and N.G.H.

Competing interests:

There are no competing interests to declare.

Table 1 The research studies from which we draw data and the course and institution types represented by these data. Values in the rightmost column represent the total number of each type represented in the final predictions.

		[11]	[28]	[26]	Current study	Total
Total classes		2	29	26	12	69
Highest degree granted at institution	Associate's	0	0	1	0	1
	Bachelor's	0	0	1	0	1
	Master's	0	0	9	0	9
	PhD	2	29	15	12	58
Research designations of institution	R1	0	29	8	12	49
	R2	0	0	5	0	5
	RCU	2	0	4	0	6
	None	0	0	9	0	9
Public/private institutions	Public	2	29	16	0	47
	Private	0	0	10	12	22
Disciplines	Physics	0	0	24	10	34
	Astronomy	0	0	2	2	4
	Biology	2	29	0	0	31
Class sizes	Less than 50	0	3	15	1	19
	50-150	0	3	6	3	12
	More than 150	2	23	5	8	38
Class levels	Lower division	2	22	26	12	62
	Upper division	0	7	0	0	7

Data and materials availability:

All data and source code are available at https://doi.org/10.5281/zenodo.19598369

References

[1] Aiken JM, De Bin R, Lewandowski HJ, et al (2021) Framework for evaluating statistical models in physics education research. Physical Review Physics Education Research 17(2):020104. https://doi.org/10.1103/PhysRevPhysEducRes.17.020104

[2] Andrews TM, Leonard MJ, Colgrove CA, et al (2011) Active learning not associated with student learning in a random sample of college biology courses. CBE–Life Sciences Education 10(4):394–405. https://doi.org/10.1187/cbe.11-07-0061

[3] Bazett T, Clough CL (2021) Course coordination as an avenue to departmental culture change. PRIMUS 31(3-5):467–482. https://doi.org/10.1080/10511970.2020.1793853

[4] Beichner RJ, Saul JM, Abbott DS, et al (2007) The student-centered activities for large enrollment undergraduate programs (scale-up) project. Reviews in Physics Education Research 1(1):1–42. https://doi.org/10.1119/RevPERv1.1.4

[5] Bojinova E, Oigara J (2013) Teaching and learning with clickers in higher education. International Journal of Teaching and Learning in Higher Education 25(2):154–165

[6] Burkholder E, Walsh C, Holmes NG (2020) Examination of quantitative methods for analyzing data from concept inventories. Physical Review Physics Education Research 16(1):010141. https://doi.org/10.1103/PhysRevPhysEducRes.16.010141

[7] Chi MTH (2009) Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science 1(1):73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

[8] Chi MTH, Wylie R (2014) The icap framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist 49(4):219–243. https://doi.org/10.1080/00461520.2014.965823

[9] Chin C, Osborne J (2008) Students' questions: A potential resource for teaching and learning science. Studies in Science Education 44(1):1–39. https://doi.org/10.1080/03057260701828101

[10] Commeford K, Brewe E, Traxler A (2022) Characterizing active learning environments in physics using latent profile analysis. Physical Review Physics Education Research 18(1):010113. https://doi.org/10.1103/PhysRevPhysEducRes.18.010113,

[11] Connell GL, Donovan DA, Chambers TG (2016) Increasing the use of student-centered pedagogies from moderate to high improves student learning and

attitudes about biology. CBE–Life Sciences Education 15(1):ar3. https://doi.org/10.1187/cbe.15-03-0062

[12] Crouch CH, Mazur E (2001) Peer instruction: Ten years of experience and results. American Journal of Physics 69(9):970–977. https://doi.org/10.1119/1.1374249

[13] Dancy M, Henderson C, Apkarian N, et al (2024) Physics instructors' knowledge and use of active learning has increased over the last decade but most still lecture too much. Physical Review Physics Education Research 20(1):010119. https://doi.org/10.1103/PhysRevPhysEducRes.20.010119

[14] Freeman S, Eddy SL, McDonough M, et al (2014) Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences 111(23):8410–8415. https://doi.org/10.1073/pnas.1319030111

[15] Lombardi D, Shipley TF (2021) The curious construct of active learning. Psychological Science in the Public Interest 22(1):8–43. https://doi.org/10.1177/1529100620973974

[16] Lund TJ, Pilarz M, Velasco JB, et al (2015) The best of both worlds: Building on the copus and rtop observation protocols to easily and reliably measure various levels of reformed instructional practice. CBE–Life Sciences Education 14(2):ar18. https://doi.org/10.1187/cbe.14-10-0168

[17] Nissen J, Donatello R, Van Dusen B (2019) Missing data and bias in physics education research: A case for using multiple imputation. Physical Review Physics Education Research 15(2):020106. https://doi.org/10.1103/PhysRevPhysEducRes.15.020106

[18] Nissen JM, Talbot RM, Nasim Thompson A, et al (2018) Comparison of normalized gain and cohen's d for analyzing gains on concept inventories. Physical Review Physics Education Research 14(1):010115. https://doi.org/10.1103/PhysRevPhysEducRes.14.010115

[19] Olson S, Riordan DG (2012) Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics

[20] Prather EE, Rudolph AL, Brissenden G, et al (2009) A national study assessing the teaching and learning of introductory astronomy. part i. the effect of interactive instruction. American Journal of Physics 77(4):320–330. https://doi.org/10.1119/1.3065023

[21] Rowe MB (1986) Wait time: Slowing down may be a way of speeding up! Journal of Teacher Education 37(1):43–50. https://doi.org/10.1177/002248718603700110

[22] Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with python. SciPy Conference Proceedings https://doi.org/10.25080/Majora-92bf1922-011

[23] Shmueli G (2010) To explain or to predict? Statistical Science 25(3):289–310. https://doi.org/10.1214/10-STS330

[24] Smith MK, Jones FHM, Gilbert SL, et al (2013) The classroom observation protocol for undergraduate stem (copus): A new instrument to characterize university stem classroom practices. CBE–Life Sciences Education 12(4):618–627. https://doi.org/10.1187/cbe.13-08-0154

[25] Stains M, Harshman J, Barker MK, et al (2018) Anatomy of stem teaching in american universities: A snapshot from a large-scale observation study. Science 359(6383):1468–1470. https://doi.org/10.1126/science.aap8892

[26] Sundstrom M, Gambrell J, Green C, et al (2026) Relative benefits of different active learning methods to conceptual physics learning. Nature Physics pp 1–6. https://doi.org/10.1038/s41567-026-03307-2,

[27] Theobald EJ, Hill MJ, Tran E, et al (2020) Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. Proceedings of the National Academy of Sciences 117(12):6476–6483. https://doi.org/10.1073/pnas.1916903117,

[28] Weir LK, Barker MK, McDonnell LM, et al (2019) Small changes, big gains: A curriculum-wide study of teaching practices and student learning in undergraduate biology. PLOS ONE 14(8):e0220900. https://doi.org/10.1371/journal.pone.0220900