Applying model analysis to a resource-based analysis of the Force and Motion Conceptual Evaluation

Trevor I. Smith,¹ Michael C. Wittmann,² and Tom Carter³

¹*Department of Physics and Astronomy, Dickinson College, Carlisle, Pennsylvania 17013, USA*

²*Department of Physics and Astronomy, and College of Education and Human Development, University of Maine, Orono, Maine 04469, USA*

³*Natural and Applied Sciences Division, College of DuPage, Glen Elyn, Illinois 60137, USA*

(Received 19 April 2013; published 2 July 2014)

Previously, we analyzed the Force and Motion Conceptual Evaluation in terms of a resources-based model that allows for clustering of questions so as to provide useful information on how students correctly or incorrectly reason about physics. In this paper, we apply model analysis to show that the associated model plots provide more information regarding the results of investigations using these question clusters than normalized gain graphs. We provide examples from two different institutions to show how the use of model analysis with our redefined clusters can provide previously hidden insight into the effectiveness of instruction.

DOI: [10.1103/PhysRevSTPER.10.020102](https://doi.org/10.1103/PhysRevSTPER.10.020102)

PACS numbers: 01.40.Fk, 01.40.gf, 01.40.G-

I. INTRODUCTION

In this paper, we apply model analysis [1,2] to data gathered using the Force and Motion Conceptual Evaluation (FMCE) [3] in which questions have been clustered according to a resources-based analysis of students' correct and common incorrect responses [4]. The standard practice with the FMCE, borrowed from the methods used with the Force Concept Inventory (FCI), has been to administer the test at the beginning and end of a course, calculate the average normalized gain ($\langle g \rangle$) of the class, and use statistical analyses to compare different classes [3,5–8]. At times, subsections of the FCI and FMCE are analyzed according to the normalized gain of just those questions [9,10]. There are debates, though, about the proper choice of question clusters [4] as well as the validity of using the normalized gain on the FMCE as a whole, much less on subsections of the FMCE [8]. To address the former, we have previously presented a revision to these practices that includes a redefinition of how questions are grouped into clusters [4], paying particular attention to the ideas with which students enter a course and the insight that can be gained in interpreting students' incorrect responses to the questions on the FMCE. To address the latter, we discuss in this paper an alternative approach that does not rely on normalized gain and provides a different representation of a class's performance on the FMCE.

The data we present come from two institutions that have carried out many different reform activities in their algebra-based physics courses. One is a land-grant research-intensive institution, the other a two-year college. The emphasis of this paper is not on comparisons between the schools nor on evaluation of the many implemented course reforms. Instead, our paper emphasizes the new form of analysis, and we place information about the course reforms at each institution in the appendixes for possible future comparison.

The analysis in this paper leads us to recommend the use of our previously defined, resource-based question clusters as the basis elements of model analysis plots [2]. We first present data from two institutions on specific clusters of questions on the FMCE. We begin by using an old representation, namely, histograms of normalized gains combined with statistical analyses. We then summarize the model analysis representation and apply it to these same data. Representing the data in this way allows us to answer questions that cannot be resolved by considering normalized gain alone.

II. TWO WAYS OF ANALYZING FMCE DATA

In this section, we describe the two methods we have used to analyze data from the FMCE. In both cases, we make use of question clusters defined by common incorrect responses as determined by a resources-based analysis (see Table I for question cluster definitions) [4]. Our first form of analysis is to calculate normalized gains on the FMCE for the entire test and each of the different question clusters. Our second method is to carry out model analysis. We review each of these methods in this section before comparing data in the next section.

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

TABLE I. Revised questions clusters on the FMCE with corresponding correct and common incorrect models. See Ref. [4] for full details.

Cluster	Questions	Correct model	Incorrect model(s)
Force Sled	1–4, 7	force proportional to acceleration	force proportional to velocity
Reversing Direction	8–13, 27–29	constant downward force or acceleration	force or acceleration in the direction of motion
Force Graphs	14, 16–21	force proportional to acceleration	force proportional to velocity
Acceleration Graphs	22–26	acceleration proportional to the change in velocity	acceleration proportional to velocity
Newton III	30–32, 34, 36, 38	objects exerts equal forces on each other	(a) mass dependence: more massive objects exert more force; (b) action dependence: more active objects exert more force (i.e., those that are faster or responsible for the motion of another object when pushing)
Velocity Graphs	40–43	velocity proportional to the change in position	velocity proportional to position
Energy	44–47	kinetic energy or velocity at the bottom of a hill depends on the height of the hill	kinetic energy or velocity at the bottom of a hill depends on the slope of the hill

A. Normalized gains

The normalized gain reports a class's improvement as a fraction of the possible improvement that the class could make over the course. Two different methods can be used to calculate the gain. One can look at the normalized gain of the average pre- and postinstruction scores of a class:

$$\langle g \rangle = \frac{\langle \text{post} \rangle - \langle \text{pre} \rangle}{\text{max} - \langle \text{pre} \rangle}, \quad (1)$$

or one can look at each individual student's normalized gain on the test:

$$g = \frac{\text{post} - \text{pre}}{\text{max} - \text{pre}}, \quad (2)$$

and then take the average of the individual gains ($\bar{g}$) [11]. In each of these equations, “pre” is the score obtained on the preinstruction assessment, “post” is the score obtained on the postinstruction assessment, and “max” is the maximum possible score. As an example, a class with a pretest average of 20% and a post-test average of 60% improves by 40%, which is half of the possible improvement of 80%, giving a normalized gain of 0.5. Similarly, a class moving from 80% to 90% has a normalized gain of 0.5, but a class making the same 10% gain from 20% to 30% has a normalized gain of only 0.125. Hake introduced the measure to the physics education research community, and his results and those summarized by others indicate that classes with similar forms of instruction typically have similar normalized gains on the FCI [5,7,12].

To determine whether a difference in normalized gains between two (or more) populations is significant, we rely on statistical analyses. One common method is to perform a one-way analysis of variance (ANOVA) on the set of

students' individual normalized gains. The resulting p value indicates whether or not statistically significant differences exist. For our analyses we have set the threshold for significance at $p \leq 0.05$. When using ANOVA to compare three or more populations, the main effect determines whether a statistically significant difference exists between any of the populations. *Post hoc* pairwise comparisons show more precisely between which groups the difference occurs and which group's score is higher. If the main effect is not found to be statistically significant (i.e., $p > 0.05$), it is inappropriate to claim that a difference exists between the groups (even if a pairwise comparison yields a statistically significant result) [13].

Using a one-way ANOVA requires that individual student normalized gains be calculated. However, an individual student's pre- and post-test scores may not allow for a meaningful calculation of g using Eq. (2). For example, a student who was completely correct both before and after instruction would have $g = \frac{0}{0}$ (since $\text{pre} = \text{post} = \text{max}$). This could be interpreted as $g = 0$, as $g = 1$, or as an undefined quantity; in any case, it does not provide a meaningful measure of how much this student learned during the course [14]. Additionally, a student may score lower on the post-test than on the pretest. In this case we replace normalized gain with normalized loss,

$$l = \frac{\text{post} - \text{pre}}{\text{pre}}, \quad (3)$$

so that the “normalized change” (gain or loss) is always within the interval $[-1, 1]$ [15,16]. In practice these cases for which Eq. (2) does not yield a meaningful result are rare when looking at the FMCE as a whole, but may be more prevalent when looking at a single cluster of questions [6,16]. In this paper, we present an example of such data in

the context of FMCE questions concerning velocity (see Sec. III A 4). Of paramount importance for calculating g is the requirement that all data be matched pre- and post-instruction. Without matched data, g cannot be calculated; therefore, some data may be omitted from statistical analyses.

A second option for determining the statistical significance of differences between populations is to seek data sets with averages that differ by more than one standard error. The most straightforward way to determine the standard error of a data set would be to calculate all students' individual gains and calculate the standard error as

$$SE = \frac{s}{\sqrt{n}}, \quad (4)$$

where s is the standard deviation of the data set of students' individual g values, and n is the number of students in the data set. However, this method still requires the calculation of individual student gains and is thus prone to the same difficulties as the one-way ANOVA.

In an effort to alleviate these concerns, we have defined a modified expression for standard error that incorporates the standard deviations of the pre- and post-test data separately. When combining two different data sets, it is common statistical practice to discuss the variance across both sets via the pooled standard deviation,

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \quad (5)$$

where n_i is the number of data points in and s_i is the standard deviation of the i th data set (p. 288 of Ref. [13]). The pooled standard error is then defined as

$$SE_{\text{pooled}} = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}} \quad (6)$$

$$= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}. \quad (7)$$

In our case, we wish to determine the standard error of the average normalized gain $\langle g \rangle$ for a class; thus, we define the pooled standard error in terms of the standard deviations of both the pre- and post-test data sets:

$$SE_{\langle g \rangle} = \frac{1}{100} \sqrt{\frac{(n_{\text{pre}} - 1)s_{\text{pre}}^2 + (n_{\text{post}} - 1)s_{\text{post}}^2}{n_{\text{pre}} + n_{\text{post}} - 2}} \times \left(\frac{1}{n_{\text{pre}}} + \frac{1}{n_{\text{post}}} \right)^{1/2}, \quad (8)$$

where the leading term ($1/100$) is needed to convert from the standard deviations of each data set (wherein scores are reported as percentages ranging from 0 to 100) to a standard error of normalized gain (which is reported as a decimal ranging from 0 to 1) [17]. The value of $SE_{\langle g \rangle}$ is different for each set of data, and $n_{\text{pre}} = n_{\text{post}}$ for matched data sets. Using the definition in Eq. (8) allows us to determine statistically significant differences without requiring matched data sets and without worrying about special cases for which Eqs. (1) and (2) give meaningless results since the chances of either $\langle \text{pre} \rangle > \langle \text{post} \rangle$ or $\langle \text{pre} \rangle = \text{max}$ are vanishingly small for class averaged data.

The average normalized gain ($\langle g \rangle$) and standard error ($SE_{\langle g \rangle}$) can be calculated for an entire data set and also for subsets of questions. In our analysis we group the questions on the FMCE according to the clusters we defined previously [4]. For matched data sets we also calculate each student's individual normalized gain (or loss) for the entire test and each question cluster and use a one-way ANOVA to determine statistically significant differences between the years for each course. As we report in Sec. III, determining statistically significant differences using ANOVA and $SE_{\langle g \rangle}$ yield nearly identical results (exceptions are discussed below).

As is common practice, we report graphs of $\langle g \rangle$ with accompanying error bars representing $SE_{\langle g \rangle}$ (see, for example, Figs. 2 and 5). However, these graphs leave out certain information. Because they only show gains in correct responses, we do not know what common incorrect responses students give. We also cannot compare the percentage of responses that are correct to the percentage of answers that are the most common incorrect responses. Because normalized gain plots do not show the actual pre- or post-test score, we do not know how much students understand (on average) at the start of the term nor whether they end with a score that shows mastery of the material. Model analysis addresses each of these shortcomings while keeping the useful information provided by the normalized gain graphs.

B. Model analysis

1. Model analysis and the model plot

As instructors and researchers, we often wish to consider more than the correctness of the students' answers on the FMCE and their normalized gains. We want to know *how* they are answering the questions and what (if anything) these answers might indicate about their understanding. Where a normalized gain graph leaves out this information, model analysis includes it with additional information.

Applying model analysis to multiple-choice data provides a method for analyzing students' responses in terms of predefined mental models and graphically representing the probability that students use each of these models [2]. When applying model analysis, each available response for a multiple-choice question is categorized in one of three

ways: (1) correct Newtonian thinking, (2) corresponding with a well-known and documented incorrect student mental model, defined in our case as the misapplication of a commonly used resource leading to the choice of an incorrect distractor, or (3) not corresponding with any previously observed mental models, “other.” Each student’s responses are classified using this categorization scheme to determine the frequency of each well-defined model throughout a class of students. Data are also kept regarding the coherence with which each student uses each of the mental models to determine whether they are in a “pure model state” (using only one mental model for answering all questions within a given domain) or a “mixed model state” (using two or more models while answering questions in a given domain). The data are analyzed to create a model plot that graphically depicts the class’s knowledge state. Plotting the class’s knowledge state at the beginning and end of the course allows for a visual representation of the change in their understanding of this particular model.

An example of a model plot can be seen in Fig. 1. A point in the model 1 (or model 2) region in Fig. 1 indicates a high probability that students consistently use model 1 (model 2) when answering questions related to a particular aspect of physics. A point in the mixed model region (e.g., point B in Fig. 1) indicates that the class as a whole is in a mixed model state. This generally occurs when some students mostly employ model 1 and some students mostly employ model 2. A “mixed model class” could also be populated by many students who are in mixed model states themselves. The distinction between these populations can be observed by the distance d of the class’s data from the “probability = 1” barrier line ($P_1 + P_2 = 1$) displayed in

FIG. 1. Depiction of the various regions of the model plot and the information that can be ascertained therein for a set of questions that provide responses that can be classified as either correct (model 1) or corresponding to a single incorrect mental model (model 2). This figure was recreated from Ref. [2], where σ_μ^2 is the μ th eigenvalue of the class model density matrix and $v_{i,\mu}$ is the i th component of the μ th eigenvector.

Fig. 1. Data close to the barrier ($d \ll 1$) indicate a class populated by students who are very similar to each other and (typically) exist in mixed model states individually; data closer to the origin indicate a class populated by students who are less similar to each other and may exist in different pure model states. Data in the secondary region (bounded by $P_1 + P_2 = 0.4$) may also indicate that aspects of student responses are not being represented and a third model must be defined [18]. Throughout our discussion we will consider a class of students to be *consistent* if they tend to provide the same answers as their classmates (as indicated by a point near the $P_1 + P_2 = 1$ barrier: $d \ll 1$). We will consider students to be *coherent* if they tend to individually provide answers that subscribe to a single model (as indicated by a point in either the model 1 or model 2 region or far from the barrier line in the mixed region).

2. Adding errors bars to the model plot

One shortcoming of model analysis as presented in Ref. [2] is the lack of error bars on the model plot representation or other measures of statistical uncertainty. Sommer and Lindell recognized this omission and proposed a method for determining the uncertainty in the eigenvalues of the class density matrix [19]. Their method considers that the measured probability that a student uses a particular model p_i may have an associated uncertainty ϵ_i such that the real probability is within the range $p_i \pm \epsilon_i$. This results in a class density matrix D ,

$$D = \begin{pmatrix} p_1 & \sqrt{p_1 p_2} & \cdots & \sqrt{p_1 p_n} \\ \sqrt{p_2 p_1} & p_2 & \cdots & \sqrt{p_2 p_n} \\ \vdots & & \ddots & \vdots \\ \sqrt{p_n p_1} & \sqrt{p_n p_2} & \cdots & p_n \end{pmatrix}, \quad (9)$$

and an associated general error matrix E ,

$$E = \begin{pmatrix} \epsilon_1 & E_{12} & \cdots & E_{1n} \\ E_{21} & \epsilon_2 & \cdots & E_{2n} \\ \vdots & & \ddots & \vdots \\ E_{n1} & E_{n2} & \cdots & \epsilon_n \end{pmatrix}, \quad (10)$$

where

$$E_{ij} \equiv \sqrt{p_i p_j + p_i \epsilon_j + p_j \epsilon_i + \epsilon_i \epsilon_j} - \sqrt{p_i p_j}. \quad (11)$$

Given that the error in the measured probability could be positive or negative, each term in E could also be either positive or negative. This information is used to generate a set of specific error matrices. Because the $(n \times n)$ density matrix is symmetric, the general error matrix is also

symmetric, yielding $2^{n(n+1)/2}$ specific matrices with different combinations of positive and negative terms. By adding each of these specific error matrices to the class density matrix D and computing the eigenvalues of each of the resulting matrices, one can determine the upper and lower bounds for each of the eigenvalues [19]. We may now be confident that the actual eigenvalue falls within the range $\sigma_\mu^2 \pm \Delta_\mu$, where the uncertainty Δ_μ is defined by the upper and lower bounds.

While this is a step in the right direction, it falls short of providing a mechanism for representing statistical uncertainty within the model plot (the points on which depend on both eigenvalues and the associated eigenvectors). Moreover, this method requires an initial assumption of the values of the uncertainties ϵ_i that are used to create the general error matrix. Sommer and Lindell propose using a single uncertainty for simplicity ($\epsilon = \max\{\epsilon_1, \epsilon_2, \dots, \epsilon_n\}$) but provide no straightforward method for determining an initial estimate. To implement their model, we use the standard error as calculated by Eq. (8) as an initial estimate of the uncertainty ϵ and calculate the uncertainty of the eigenvalue Δ_μ as the average of the difference between the upper and lower bounds and the measured eigenvalue,

$$\Delta_\mu = \frac{(\sigma_{\mu,\max}^2 - \sigma_\mu^2) + (\sigma_\mu^2 - \sigma_{\mu,\min}^2)}{2}. \quad (12)$$

We also expand their method by considering the uncertainty in the eigenvalue as a percentage and use the same percent uncertainty for each dimension on the model plot, allowing for the creation of error bars.

III. COMPARING NORMALIZED GAINS AND MODEL ANALYSIS

Data for this study come from two very different sources: A four-year research-intensive land-grant university (school 1) and a two-year college (school 2). Data were gathered from the general physics course over several years at each institution using standard practice: students answered the FMCE at the beginning of the semester (before all instruction) and again at the end of the semester (after all instruction), and no credit was given for correctness. At each school the general physics course is a year-long algebra-based introductory course that employs various interactive-engagement instructional techniques (e.g., University of Washington-style tutorials, interactive lecture demonstrations, etc.). The FMCE was administered at the beginning and end of the portion of the course that covered mechanics. During the years in which data were gathered, several modifications were made to the instructional strategies in each course. As our focus is the comparison between reports of normalized gain and model analysis plots, we consider the results from each school separately and defer the presentation of specific curricular changes to the Appendix A. We would like the reader to

consider the differences between graphs of normalized gain and model analysis plots (in terms of the information that is conveyed in each) before being concerned with why the results differ from year to year.

A. Comparing data at school 1

The general physics course at school 1 is a two-semester algebra-based introductory sequence that covers linear and rotational kinematics and dynamics, work and energy, gravitation, waves and sound, electrostatics, circuits, magnetostatics, electrodynamics, and optics. Data for this study were gathered at the beginning and the end of the first semester, by which time the students had completed their study of classical mechanics. The course is designed with two hours of lecture, two hours of recitation (or tutorial), and one two-hour session of laboratory work each week. Teaching assistants are responsible for running both the recitation and laboratory sections of the course. Table II shows the number of students who completed the pretest, post-test, and both for all three years. All students who completed either the pretest or the post-test were included in calculations of $\langle g \rangle$ and $SE_{\langle g \rangle}$ and in model analysis, but only matched data were included in ANOVA analyses.

Figure 2 presents the average normalized gains of the overall FMCE score in all three years at school 1. Figure 3 shows the $\langle g \rangle$ for all question clusters in each of these years. The error bars in the figures represent the standard error of the gains calculated using Eq. (8). Results of a one-way ANOVA show that the main effect between years is only significant for three question clusters at school 1. The results of *Post hoc* comparisons between instructional years for these clusters are displayed in Table III.

Comparing Table III and Fig. 3 we clearly see non-overlapping error bars in all clusters with a statistically significant main effect. We also see that clusters in which all error bars overlap do not have significant main effects, as expected. However, the Velocity Graphs and Energy clusters clearly have nonoverlapping error bars in Fig. 3 (for some years), but do not have significant main effects in Table III. Surely a normalized gain of 0.46 ± 0.04 (Velocity Graphs, year 3) is significantly different than a gain of 0.05 ± 0.05 (Velocity Graphs, year 1). But why, then, do the ANOVA results not reflect this difference? Are these years really different from each other or not? We cannot

TABLE II. Number of students in each data set at school 1. Some students in year 1 did not complete all sections of the FMCE, so the number of matched data points is lower for the Newton III (62), Velocity Graphs (61), and Energy (58) clusters.

Year	Pretest	Post-test	Matched
1	113	66	58–64
2	109	85	75
3	111	90	84

FIG. 2. FMCE normalized gain comparisons for school 1. Error bars are the standard error of the gain.

FIG. 3. Normalized gains for school 1 for each question cluster on the FMCE. Error bars are the standard error of the gain.

answer these questions using normalized gain alone and, therefore, turn to model analysis to shed additional light on these claims.

We present model analysis plots for the three question clusters showing significant main effects as well as the Velocity Graphs and Energy clusters in Fig. 4 (plots for additional question clusters are included in Appendix B). In each plot, the vertical axis is the “correct model” while the horizontal axis is a model consisting of the incorrect application of commonly used resources, as described in Ref. [4]. Because students often use two different incorrect models for reasoning about Newton’s third law [1,4,20], the Newton III cluster is depicted in two different model plots: one showing students’ use of mass dependence reasoning and one showing students’ use of action dependence reasoning. As a result, we show six model analysis plots.

We compare the information presented by the normalized gain graphs (Fig. 3) and the model analysis plots (Fig. 4) in terms of the individual question clusters. In all cases, the data point closer to the lower right corner of the model plot represents pretest data, and the point closer to the upper left corner represents post-test data. We pay particular attention to the length and slope of the line connecting the pre- and postinstruction data points for each year: longer lines indicate greater change, a slope that is steeper than -1 indicates a class that is becoming more

TABLE III. Results from one-way ANOVA and *Post hoc* comparisons between years at school 1. Numbers reported are p values with the threshold for significance set at $p \leq 0.05$. The significance values indicate that statistically significant differences do not exist within the homogeneous subsets designated as “higher” or “lower” years. Years that appear twice belong to both homogeneous subsets.

Main effect		Higher		Significance
		Lower		
Overall	<0.001	3	2	0.148
		1		1.000
Force Sled	0.001	3	2	0.177
		1	3	0.108
Reversing Direction	0.009		2	1.000
		1	3	0.917
Force Graphs	0.160	...		...
Acceleration Graphs	0.162	...		...
Newton III	<0.001	3	2	0.460
		1		1.000
Velocity Graphs	0.223	...		...
Energy	0.128	...		...

consistent, and a slope that is shallower than -1 indicates a class that is becoming less consistent.

1. Improving toward mastery: Newton III

The error bars on the model analysis plot provide the same conclusion as the normalized gain graph—that students from years 2 and 3 improved their scores on questions related to Newton’s third law more than the year 1 students—but the model analysis plot reveals more information as well. The preinstruction data points for all three years are nearly on top of each other on both the Newton III-mass and Newton III-action plots. This indicates that classes start each year with the same mix of responses, corresponding to the same likelihood of using coherent incorrect lines of reasoning when answering questions about physics (and a very small probability of using the correct model).

Before instruction, students use the mass dependence model much less frequently than the action dependence model when answering questions on the FMCE, as indicated by the horizontal coordinate on the model plot. After instruction, it is very unlikely for students in years 2 and 3 to use the mass dependence model, as shown by the very small horizontal coordinate on the Newton III-mass plot. The change in year 1 on the Newton III-mass plot is negligible, as indicated by the overlapping pre- and postinstruction error bars.

Considering the Newton III-action plot, one can see that the classes in years 2 and 3 stayed about as consistent as when they began the semester, as indicated by their connecting lines being roughly parallel to the $P = 1$ barrier. We interpret this as evidence that the class as a whole

FIG. 4. Model analysis plots of FMCE data from school 1. Normalized gain plots for each cluster are included to allow for easier comparison of representations. Plots for additional question clusters are presented in Appendix B.

learned together: most students who started in a mixed model state progressed to a pure correct state, and most students who started in a pure incorrect state progressed to a mixed state. However, the year 1 class has a connecting line that is much shallower, indicating a decrease in consistency. We interpret this as an indication that the majority of students were unaffected by instruction, but that a minority improved to either a mixed state or a pure correct state. This leaves the year 1 class still in the action dependence region of the plot while the classes in years 2 and 3 progressed to the border between a mixed model state and the correct Newtonian state.

So, while the normalized gain graph shows that the classes of years 2 and 3 learned more than the class of year 1 (with roughly 3.5 times the normalized gain), the model analysis plot gives more detail about what this difference means and that the final result is a more consistent class of students, most of whom are using the correct Newtonian model to one degree or another. The model analysis plots also tell us that the students in all three years began the semester with the same level of understanding of Newton's third law (as indicated by their responses to the FMCE).

2. Less consistent: Reversing Direction

In the Reversing Direction cluster, normalized gains are modest, but there is still a significant difference between students in year 2 compared to those from years 1 and 3 (as indicated by both the error bars in Fig. 3 and the reported p values in Table III).

The model analysis plot again provides more information. As with the Newton III plots, we see that classes start the semester at essentially the same location on the plot, again indicating remarkable consistency in their preinstruction model use. The error bars on the model plot indicate that classes in years 1 and 3 end the semester with the same level of understanding, and that the class in year 2 is statistically different (as expected by the normalized gain graphs and ANOVA results). However, the change in student reasoning is such that students are still in the model 2 region of the plot at the end of each semester; their use of the correct model has not improved substantially.

As seen with year 1 in the Newton III data, the change in all three years is mostly horizontal, indicating a decrease in the consistency of the students in each class; i.e., the students within each class are not answering the same as

each other. One interpretation is that mentioned above: that a minority of students are improving toward the correct model (constant downward force or acceleration), but that a majority are unaffected by instruction. However, we see a dramatic decrease in the use of the most common incorrect model (force or acceleration in the direction of motion) without much increase in the use of the correct model. A more likely interpretation of this result is that students in these classes (particularly years 1 and 3) end the semester no longer thinking what we do not want them to think, but not yet thinking what we want them to think. That is, a different incorrect model is emerging and students are providing answers that are more consistent with the “other” model. This interpretation is supported by the shift toward the secondary region of the model plot. In fact, if a third dimension was shown to represent the “other” model, the year 1 class would move from a coordinate of 0.03 before instruction to a coordinate of 0.11 after instruction.

The statistically significant result shown in Fig. 3 and Table III can be seen in the Reversing Direction model plot by the fact that the year 2 postinstruction data point is higher than that of either year 1 or year 3. However, Fig. 4 also shows us that, while statistically significant, this result is not very impressive, as the probability of students using the correct model at the end of the semester is quite low for all classes. The important message that is not shown by the normalized gain graphs is that the students *are* being affected by instruction, just not in the way that the instructor likely intended.

3. Mixed improvement: Force Sled

In the Newton III question cluster, some classes improved significantly and arrived at a relatively consistent use of the correct physics. In the Reversing Direction question cluster, there were significant differences in the normalized gain, but no real improvement in the classes’ use of the correct model. The Force Sled question cluster rests between these two extremes: classes show some improvement but do not end the semester with coherent and consistently correct responses.

The $SE_{(g)}$ error bars on the normalized gain graph show that year 2 had a significantly larger gain than both years 1 and 3, and that year 3 was larger than year 1, but not quite significant. The ANOVA results in Table III show a gain in year 2 that is statistically higher than year 1, but not significantly higher than year 3. ANOVA results agree that years 1 and 3 are not significantly different. Moreover, the Force Sled normalized gain graph looks fairly similar to that of the Reversing Direction cluster, as year 2 has $\langle g \rangle \approx 0.32$ and years 1 and 3 both have $\langle g \rangle < 0.2$. Is student understanding of the topics assessed by these two clusters linked in some way? Is the year 2 class different from the year 3 class or not? To answer these questions, consider the Force Sled model plot in Fig. 4.

The Force Sled plot in Fig. 4 reveals that the classes from years 1 and 3 remain in the model 2 region throughout the semester, though they improve slightly. The year 2 class ends the semester barely in the mixed model region, which does not indicate successful learning of the desired correct model, though it does show the greatest improvement across the three classes. The error bars on the model plot suggest that all three classes begin the semester with the same likelihood of using the correct or common incorrect model, and that these classes end the semester at three different levels of understanding.

As with the Reversing Direction question cluster, the slopes of the connecting lines on the Force Sled plot indicate that the students in each year are becoming less consistent with each other as the semester progresses. However, the severity of this phenomenon is not as great as in the Reversing Direction cluster and may predominantly result from some students developing a greater understanding of the material more quickly than their classmates. From Fig. 4 one can certainly make the argument that student responses to the Force Sled questions were more correct after instruction than their responses to the Reversing Directions questions (particularly in year 2). This difference is completely hidden in Fig. 3 where the results from these two clusters appear quite similar. The model analysis results suggest that (a) the year 2 class performed significantly better than both year 1 and year 3, and (b) student understanding of concepts related to the Force Sled and Reversing Direction cluster are not directly linked.

4. Statistical discrepancies: Velocity Graphs and Energy

Figure 3 seems to indicate that students in year 3 had statistically significantly higher gains on the Velocity Graphs question cluster and significantly lower gains on the Energy question cluster than those in either of the previous two years. The results displayed in Table III, however, indicate a main effect that is not statistically significant for either cluster ($p = 0.14$ and 0.19 , respectively). The Velocity Graphs and Energy plots in Fig. 4 provide more information to help interpret these results.

From the Velocity Graphs model plot we see that the majority of students in all three years started the semester answering nearly all of the Velocity Graphs questions correctly, as indicated by the preinstruction data points well within the model 1 region. We also see that the variation in both the preinstruction and postinstruction scores is small, and that the error bars of all points (pre- and postinstruction) overlap each other. However, variation does exist between the years.

Figure 4 makes it clear that the statistically significant difference depicted in Fig. 3 is due to slightly lower preinstruction scores combined with slightly higher postinstruction scores in year 3 rather than any significant

difference in pre- or postinstruction understanding. We feel this is an excellent example of the ceiling effect: when the denominator in Eq. (1) is small, small variations in the numerator yield large differences in $\langle g \rangle$. In addition to the ceiling effect, we point to the distinction between the gain of the averages ($\langle g \rangle$) and the average of the individual gains ($\bar{g}$) [15]. We find that some students “trade” scores: e.g., some went from 100% to 75% correct and others went from 75% to 100% correct. A pair of students who trade scores in this manner contribute nothing to the gain of the averages, but they have an average individual gain of 0.375. When combined, the differences between the calculations of $\langle g \rangle$ and $\bar{g}$ account for the anomaly between the ANOVA results and the interpretation of the error bars in Fig. 3.

From the Energy model analysis plot we see that students in years 1 and 2 end the course in the model 1 region, while the year 3 class ends in the mixed model region. Furthermore, error bars on the pretest data points show that all three classes started the semester with the same level of understanding. Therefore, it seems that even though a one-way ANOVA did not yield statistically significant results, a meaningful difference exists between year 3 and years 1 and 2 in terms of students’ understanding of work and energy concepts at the end of the course. This is another case in which the difference between $\langle g \rangle$ and $\bar{g}$ caused discrepant results between normalized gain graphs and ANOVA results, but the model analysis results allow us to view the data in a different way that highlights the similarities and differences between the classes in each year based on their use of the correct and common incorrect models.

5. Summary of results

We observe that some classes did better than others at school 1; typically the year 2 class shows the greatest improvement compared to years 1 and 3. More importantly, we observe that the model analysis plots let us make claims about *what kinds of change* occurred. Questions we can now ask and answer include the following: Did the final class performance indicate consistent mastery of the topic, or were classes improving but not enough? Were students in a particular class consistent with each other, and how does this consistency change over the course? How likely were students to use the most common incorrect model, and did they use other models as well? The answers to these are represented in the model analysis plots, not the normalized gain graphs.

Moreover, we find that model analysis plots may reveal differences between data sets that were not apparent from statistical analyses. In particular, we find model plots to be extremely helpful when interpreting data from the Velocity Graphs and Energy clusters for which our two normalized gain analyses provided disparate results. By placing our data within a model of learning, such as

TABLE IV. Number of students in each data set at school 2. Some students in years 2a, 2b, 4, and 7 did not complete all sections of the FMCE, so the number of matched data points is lower for some clusters.

Year	Pretest	Post-test	Matched
1a	33	13	13
1b	30	14	12
2a	35	24	20–22
2b	33	29	20–28
3	50	40	40
4	45	28	23–26
5a	46	27	26
5b	0	24	0
6	0	24	0
7	36	28	25–26

resources, and utilizing model analysis to analyze the data within this framework, we find ourselves in a position to determine whether or not these data are pedagogically and intellectually significant as well as statistically significant.

B. Results from the school 2

The general physics course at school 2 runs for either three quarter-long sessions (years 1–5) or two semester-long sessions (years 6 and 7) and covers a variety of physics topics including linear and rotational kinematics and dynamics, thermodynamics, waves and sound, and electrostatics. The students meet for four hours of lecture or recitation and two hours of laboratory work each week. The primary instructor controls all portions of the course. There are no teaching assistants. Data were collected in ten different sections over the course of seven years (two sections each in years 1, 2, and 5) [21]. No preinstruction data were collected for two of these sections (years 5b and 6); reports of normalized gains omit these years, but the postinstruction data are reported in the model analysis plots. Justification for this reporting practice is included below. Table IV shows the number of students who

FIG. 5. FMCE normalized gain comparisons for school 2. Error bars are the standard error of the gain.

FIG. 6. Normalized gains for school 2 for each question cluster on the FMCE. Error bars are the standard error of the gain.

completed the pretest, post-test, and both for all years. All students who completed either the pretest or the post-test were included in calculations of $\langle g \rangle$ and $SE_{\langle g \rangle}$ and in model analysis, but only matched data were included in ANOVA analyses.

Figure 5 presents the average normalized gains of the overall FMCE score for the eight sections in which both pre- and postinstruction data exist. Figure 6 shows the average normalized gain for all question clusters in each of these years [22]. As in Figs. 2 and 3, the error bars represent the standard error of the gains.

Using one-way ANOVA, with the threshold for significance set at $p \leq 0.05$, we find that statistically significant main effects exist for the Reversing Direction, Force Graphs, and Newton III clusters as well as on the overall FMCE. *Post hoc* analyses revealed homogeneous subsets of years with higher or lower gains. Table V shows the

results from these analyses and indicates the p value of the main effect, the years that are statistically similar and have higher scores than the others, the years that are statistically similar and have lower scores than the others, and the significance values that show that statistically significant differences do not exist within each of these groups. For each cluster, years are ordered from left to right by increasing average gains ($\bar{g}$). Years are statistically different if they are not within the same homogeneous subset.

Many of the results in Table V agree with Fig. 6: Year 7 is always in the lower subset (and isolated on the Force Graphs cluster), year 4 is always in the higher subset, years 1a and 1b are particularly low on the Newton III cluster, and all years are similar on the Acceleration Graphs and Velocity Graphs clusters. However, some discrepancies exist between the two representations of the data. Figure 6 indicates that year 7 is significantly lower than all other years on the Force Sled and Reversing Direction clusters, but the ANOVA results indicate that all years are statistically similar on the Force Sled cluster and that year 7 is only significantly different from year 4 on the Reversing Direction cluster. The significance levels on the Reversing Direction cluster suggest that the years within the “lower years” homogeneous subset may not be as similar to each other as the years in the “higher years” subset, but this is still not a statistically significant difference. Years 1a and 1b appear to be significantly lower than year 3 on the Newton III cluster in Fig. 6, but the ANOVA results place them in the same homogeneous subset (with year 7). Moreover, year 1b is grouped in the same subset as years 2a, 2b, 4, and 5a, which appear significantly higher in Fig. 6. As with school 1, we will look to the model analysis plots for additional information and clarity.

Postinstruction data from school 2 for years 5b and 6 have been included in the model plots. We feel that displaying these data allow postinstruction comparisons to be made with other years under the assumption that the students’ preinstruction knowledge state would be similar to students in other years. Figure 7 gives credence to this

TABLE V. Results from one-way ANOVA and *Post hoc* comparisons between years at School 2. Numbers reported are p -values with the threshold for significance set at $p \leq 0.05$. The significance values indicate that statistically significant differences do not exist within the homogeneous subsets designated as “higher years” or “lower years.” Years that appear twice belong to both homogeneous subsets.

Main effect		Lower years	Higher years	Significance
Overall	0.009	1a 3 1b 5a 2a 2b 4	7	0.896
		7	1a 3 1b 5a 2a	0.092
Force Sled	0.171	...	...	...
Reversing	0.026	1a 3 2a 2b 5a 1b 4	7	0.566
direction		7	1a 3 2a 2b 5a 1b	0.087
Force Graphs	<0.001	2b 3 2a 5a 1a 4 1b	7	0.739
		7	...	1.000
Acceleration	0.600	...	...	...
Graphs		...	...	...
Newton III	0.004	1b 7 3 5a 4 2a 2b	1a 1b 7 3	0.130
		1a 1b 7 3	...	0.085
Velocity Graphs	0.739	...	...	...
Energy	0.289	...	...	...

assumption as the preinstruction data points are statistically similar for all question clusters in all years. Model analysis plots for additional question clusters are included in Appendix B.

1. Mastery: Force Graphs and Force Sled

According to Fig. 6, students' responses to the questions in the Force Graphs cluster improved greatly in nearly all years of instruction, with the lowest improvement happening in year 7. These are strong results, only supported more strongly by the Force Graphs plot in Fig. 7. The post-instruction scores indicate that all classes from years 1–5a ended the semester in the model 1 region, indicating consistent and coherent mastery of the topic. The class from year 7, on the other hand, ended in the mixed model region with very little mastery. Their postinstruction data indicate that those students only began to shed their previously held beliefs in favor of correct Newtonian thinking (displaying a use of the correct model in less than 20% of their responses). Furthermore, the students in year 7 became less consistent throughout the semester, as indicated by a connecting line with a shallow slope. This supports an interpretation that some students in year 7 progressed toward mastery while their classmates did not change much over the course of the semester.

A similar pattern exists for the Force Sled cluster. The classes from years 1 to 5a ended the semester in the model 1 region, but the year 7 class never left the model 2 region. In fact, the postinstruction data point for year 7 is statistically indistinguishable from the preinstruction data points. It is clear from this representation of the data that a difference *does* exist between student performance on the Force Sled questions in year 7 and student performance in years 1–5a. It is difficult to determine why the ANOVA results do not reflect this difference, but it may be due to the difference between calculating the normalized gain of the average pre- and postinstruction class scores (as was done to generate Fig. 6) and calculating the normalized gain for each individual student (as was done to generate Table V). As seen with school 1, the impact of this difference may be subtle but quite profound [15]. From the model analysis plots we see that students from all years gained slightly greater mastery of the content assessed by the Force Graphs cluster than the Force Sled cluster. We also see that these two clusters show the same trends in the data from year to year. These trends are not observed in any of the other model analysis plots, strengthening the assertion that these clusters are closely related for these classes.

2. Mostly mastery: Newton III and Reversing Direction

According to Fig. 6, the class Newton III performance in year 1 (both sections a and b) did not improve as much

as in other years. However, the ANOVA results in Table V claim that these years are statistically similar to years 3 and 7 and that only year 1a is statistically different from years 2, 4, and 5a. The Newton III-mass and Newton III-action plots in Fig. 7 support the conclusion that student performance is significantly lower in year 1 than in later years. The error bars on the model analysis plots suggest that the two sections in year 1 are fairly similar to each other but distinct from years 2–5a and 7 [23]. We see that students begin each class with a statistically similar understanding of Newton's third law, but that the classes from year 1 did not show as large an improvement in their use of the correct Newtonian model. Where all other classes moved toward a region of clearly consistent and correct responses, the classes from year 1 moved to the mixed model region. Students in year 1 were not consistent with each other (as indicated by data points far from the $P_1 + P_2 = 1$ barrier), suggesting that some individuals in these sections may have developed a good understanding of Newton's third law while their classmates did not improve.

The Newton III model analysis plots also indicate that years 3 and 7 end the semester with an understanding of Newton's third law that is statistically similar to years 2, 4, and 5 and statistically different from year 1. The preinstruction data points in the Newton III plots in Fig. 7 show that the students from year 7 began the semester with a slightly better understanding of Newton's third law than some of the other classes. This preinstruction data combined with a slightly lower than average postinstruction score could result in a significantly lower normalized gain when the year 7 class's postinstruction understanding of Newton's third law is much closer to that of the students from years 2 to 5a than it is to the students from year 1 (as indicated by their responses to the FMCE questions regarding Newton's third law).

According to Fig. 6, the results on the Reversing Direction cluster are similar to those on the Force Sled and Force Graphs clusters. All classes from years 1 to 5a have similar normalized gains that are significantly higher than that of year 7. However, Table V indicates that year 7 is only statistically different from year 4 and that all other years are statistically similar to both year 4 and year 7. The Reversing Direction plot in Fig. 7 tells a different story in which the years are separated into three groups as defined by (non)overlapping error bars. Where the normalized gain graphs do not indicate meaningful differences between the different sections from years 1 to 5a, the model analysis plots indicate that the year 1a class and the year 3 class did not improve in the same fashion as those in years 1b, 2, 4, and 5a. For these two classes, students ended up in the mixed model region while all other classes from years 1b to 5a ended up in the model 1 region. This difference can be seen in Fig. 6 by years 1a and 3 having lower normalized gains, but they are

FIG. 7. Model analysis plots of FMCE data from school 2. Normalized gain plots for each cluster are included to allow for easier comparison of representations. Plots for additional question clusters are presented in Appendix B.

considered statistically similar to those of the other classes from years 1b to 5a. The model analysis plot in Fig. 7 clearly shows that the postinstruction understanding of Reversing Direction concepts of these two classes is different than the other four.

The year 7 class, as was seen on the model analysis plots for the Force Sled cluster, ends the semester well within the model 2 region of the Reversing Direction plot in Fig. 7, indicating very little mastery of concepts regarding situations in which the motion of an object reverses direction (having given responses indicating that the likelihood of using the correct Newtonian model was less than 10%). The Reversing Direction model plot also shows that the classes in all years became less consistent throughout the course (as indicated by data points far from the $P_1 + P_2 = 1$ barrier). As with school 1, this phenomenon may result from students beginning to use a third model for understanding the motion of and forces on objects Reversing Direction. This is especially likely in year 7 when the decrease in the probability of using the common incorrect model is greater than the increase in using the correct Newtonian model. In this way the model analysis

plot provides access to information not available when viewing normalized gains alone.

3. Reporting additional data: Years 5b and 6

One of the most beneficial aspects of using model analysis plots to display our results is the ability to include the postinstruction data from years 5b and 6 during which no preinstruction data were gathered. Without this preinstruction data, it is impossible to accurately calculate values of either average or individual students' normalized gains. Model analysis plots, however, allow us to compare the postinstruction results from these years directly with the postinstruction results from other years for which preinstruction data do exist. Including these data sets reveals additional trends.

In the Force Graphs and Force Sled plots in Fig. 7, we see that the postinstruction data points from years 5b and 6 form a kind of “bridge” from the high-gain model 1 results from years 1 to 5a and the low-gain model 2 results from year 7. On the Force Graphs plot, the year 5b class ended the course in the mixed model region, but came close to

mastering the content. Year 6 (like year 7) ended in the mixed model region, but showed very little mastery. On the Force Sled plot the students from both years 5b and 6 ended the semester in the mixed model region, with the students from year 5b closer to model 1 than those from year 6. The addition of these results seems to indicate that the transition from the high gains of years 1 to 5a and the low gain in year 7 may not have been as abrupt as it is depicted by the normalized gain graphs in Fig. 6. The model analysis plots in Fig. 7 show us that this transition happened gradually from years 5 to 7.

In the Newton III plots in Fig. 7, we see that the class from years 5b and 6 show postinstruction results very similar to those in year 7 and higher than either section in year 1. This supports our assertion that the data from year 7 may not be similar to those from year 1, but may be part of a different subset entirely. This subset (composed of years 5b, 6, and 7) appears to be similar to the results from years 2 to 5a, given the relatively large error bars.

The Reversing Direction plot in Fig. 7 shows that the postinstruction results from years 5b and 6 are similar to those in the model 1 region. Unlike the Force Sled and Force Graphs plots, we see a distinct difference between the results from years 6 and 7 on the Reversing Direction plot in which the students from year 7 had significantly lower gains and ended the semester in the model 2 region. Using model analysis plots allows us to report postinstruction data for which normalized gains do not exist. The statistically similar preinstruction data for all clusters in all years at school 2 suggest that meaningful comparisons may be made between the years using postinstruction data alone.

4. Summary of results

As with the school 1 analysis, we find that the model analysis plots in Fig. 7 provide additional information to supplement the normalized gain graphs in Fig. 6 and the ANOVA results in Table V. We observe not just the percent of the possible improvement, but whether classes were similarly consistent in their answers before instruction from year to year. We may also see whether they ended the academic term in a region that one might think of as showing mastery of the subject. Furthermore, we can observe whether students are answering consistently with each other (near the $P_1 + P_2 = 1$ barrier) or inconsistently.

Using model analysis allows us to include data sets in our analysis that would have been otherwise inadmissible (such as those from years 5b and 6). Furthermore, model analysis provides us with a “tie-breaker” of sorts when our normalized gain analyses yield contradictory results on the Force Sled cluster. The model analysis plot shows that students in year 7 ended the course with a much lower conceptual understanding of forces than those from years 1 to 5a. Along with the additional data from years 5b and

6, we are able to see a distinct declining trend in student postinstruction performance on Force Sled and Force Graphs questions from years 5 to 7 that is invisible in our reports of the normalized gain results. On the Reversing Direction cluster we are able to identify three distinct groups that are not apparent in either Fig. 6 or Table V. And we are able to use the error bars on the Newton III model plot to clarify the similarities and differences between the years.

IV. CONCLUSIONS

In this paper, we have shown that normalized gain graphs do not include all of the relevant information about a class’s learning and that model analysis plots provide a useful alternative for viewing the same data. We find that many different elements of the model analysis plots provide invaluable information. We can compare different class’s preinstruction scores to see if students entering our courses are the same from year to year. We can observe the kinds of changes made by students—are they moving toward the correct answer, or away from the incorrect one? We can also observe how the class is answering—are students gaining in mastery but still answering with mixed responses (consistent but individually incoherent, data near the $P_1 + P_2 = 1$ barrier line), or is the class as a whole gaining in mastery with some students mastering an idea and others not (individually coherent but inconsistent as a class, data far from the $P_1 + P_2 = 1$ barrier line)? We can observe how common the “common incorrect model” really is for our classes. The fundamental shift from one dimension (correct or incorrect) to a second (use of a well-defined incorrect model) increases the amount of information conveyed. With the model analysis plots portraying the beginning and ending states of a class’s understanding as well as the gain, we also include data in our analysis that would have otherwise been inadmissible (e.g., due to lack of matched preinstruction data).

We are careful not to overgeneralize the benefits of simply adding a second visual dimension to a plot representing student learning and understanding. A two-dimensional plot of normalized gains was used by Hake when he introduced the measure to physics education research [5]. By plotting gains as a function of pretest scores, Hake showed that certain regions of the plot (with common ranges of normalized gains) were filled by one type of instruction or another. This allows for the study of incoming students (are they the same every year?) but does not address the issues that the model analysis plots do. For example, we cannot determine from Hake’s plot what kinds of changes students make as a class’s score increases. Also, we cannot tell whether scores come from individual students giving incoherent answers or from students giving coherent responses but each student believing a different thing.

Thus, we believe that model analysis plots, when combined with a resources-based analysis of question clusters on the FMCE, increase the amount of detail available for understanding the kinds of learning going on in a classroom based on pre- and postinstruction survey data. However, we also recognize that histograms for comparing class-average normalized gains provide a visual clarity and a means for representing statistical significance that may be lost on model analysis plots containing multiple data points within a small visual space [24]. In this way, we feel that these representations may complement each other and that a combination of the two (as shown in Figs. 4 and 7) may be the most beneficial.

ACKNOWLEDGMENTS

We thank the Physics Education Research Laboratory at the University of Maine for insightful critiques of previous drafts. We also thank the reviewers for their comments and suggestions. This work supported in part by NSF under Grants No. DUE-0510614 and No. REC-0633954.

APPENDIX A: OUTLINE OF INSTRUCTIONAL METHODS

Data for this study come from two very different sources. We observe not only that there are large differences in class performance between the two institutions, but that the between-class performance is often different at each institution. We present brief descriptions of each institution and the curricular changes that occurred, year by year.

1. Instruction at school 1

As mentioned in Sec. III A, the general physics course at school 1 is an algebra-based introductory sequence that covers linear and rotational kinematics and dynamics, work and energy, gravitation, waves and sound, electrostatics, circuits, electrodynamics, magnetism, and optics. The course is designed with two hours of lecture, two hours of recitation (or tutorial), and one two-hour session of laboratory work each week. Teaching assistants are responsible for running both the recitation and laboratory sections of the course.

Over the span of several years school 1 made significant changes to its course structure including the use of interactive lecture demonstrations (ILDs) [26], University of Washington-style tutorials [27–29], and a modified version of the modeling method [30]. The *Tutorials in Introductory Physics* (TIP) [27] had been used in the general physics course at school 1 in either developmental or published form for several years prior to data collection, with classes using the published versions in the years for which we present data. The primary instructor in year 1 was different from the primary instructor who taught both years 2 and 3. Table VI summarizes the curricular changes that occurred during this study.

TABLE VI. Changes that were made each year to the general physics course at school 1.

Year 1	Standard lecture format <i>Tutorials in Introductory Physics</i> (TIP) [27] materials used during all “recitation” periods Confirmation-style laboratory work
Year 2	New primary instructor <i>Interactive Lecture Demonstrations</i> (ILDs) [26] implemented for some lectures Studies conducted in Newton III and energy tutorial periods in which portions of students were given tutorials other than those from the TIP materials Study conducted in lab periods in which a modified version of the modeling method [30] was utilized
Year 3	More ILDs used Several TIP tutorials replaced by alternate OST versions Modified modeling method implemented for all sections Primary instructor trained teaching assistants for tutorial and laboratory sessions

In year 2 several studies were conducted within the general physics course at school 1 to determine the effectiveness of various laboratory and tutorial materials. One study implemented a modified version of the modeling method developed by Wells, Hestenes, and Swackhamer [30] in the laboratory portion of the course for half of the student population [31]. Another investigated the effects of implementing different versions of tutorials for teaching Newton’s third law [20]. In this study one-third of the students were administered each of three tutorials: the Force: Newton’s Second and Third Laws tutorial from the TIP, the Newton 3 tutorial from the *Activity-Based Tutorials* [28], and the counterintuitive ideas: Newton’s third law tutorial from the open source tutorials (OST) [29]. A third study replaced the TIP work-energy tutorial with a locally developed tutorial for half of the students [32],[33]. Additionally, the primary instructor began using ILDs [26] coupled with the use of a personal response system (PRS) to collect student feedback.

In year 3 ILDs were used more frequently, the modified modeling method experiments were implemented within all laboratory sections, and several of the OST tutorials replaced TIP versions. The studies regarding the effectiveness of tutorials for teaching Newton’s third law and the work-energy theorem were repeated in year 3.

2. Instruction at school 2

As mentioned in Sec. III B the general physics course at school 2 covers a variety of physics topics including linear and rotational kinematics and dynamics, thermodynamics, waves and sound, and electrostatics. The students meet for four hours of lecture or recitation and two hours of laboratory work each week. The primary instructor controls

TABLE VII. Changes that were made each year to the general physics course at school 2.

Year 1	Lecture format combined with problem-solving recitations Some small-group problem-solving tasks in lab (no enforced structure) Began implementing <i>Tools for Scientific Thinking</i> (TST) labs [34] Completed approximately 10 labs
Year 2a	Replaced TST labs with <i>RealTime Physics</i> (RTP) labs [35] Increased from 10 to 14 lab sessions Reading assignments required students to answer simple questions about each chapter before the material was discussed in class Tests and examples modified to be gender neutral Small-group problem-solving tasks completed in lecture periods rather than during labs
Year 2b	Implemented <i>Peer Instruction</i> [36] in lectures using a pupil response system (PRS) Some questions modified locally
Year 3	Implemented Tycho homework system Students not allowed to drop course after halfway point
Year 4	Students allowed to drop course if they spoke with a school 2 counselor
Year 5	Lesson timing adversely affected by new text Only one small-group problem-solving session throughout the course
Year 6	Introductory course goes from 3 quarters to 2 semesters
Year 7	No prerequisite math courses No small-group problem-solving sessions Peer instruction used a small number of University of Massachusetts Physics Education Research Group questions

all portions of the course. There are no teaching assistants. We describe the changes made in lecture or problem solving and laboratory in chronological order. Table VII summarizes the curricular changes that occurred in all parts of the course at school 2 during this study.

In year 1, school 2 began implementing various pieces of research-based curricula starting with the *Tools for Scientific Thinking* (TST) laboratory materials [34]. Within the lab periods students were also occasionally required to perform problem-solving tasks within small groups that were not directly related to the current laboratory experiment.

In year 2 the TST labs were replaced by *RealTime Physics* (RTP) versions [35], and the number of laboratory experiments increased. Also in year 2 the small-group problem-solving activities were moved from the lab periods to the lecture periods. Students were required to complete reading assignments prior to lecture periods

to familiarize themselves with the basics of the material to be covered, and all tests and examples were modified to be gender neutral [37]. *Peer Instruction* developed by Mazur [36] was implemented using a PRS in year 2b but not in year 2a.

In year 3 the web-based homework and tutoring system Tycho was first used [39], and the instructor enacted a policy that prohibited students from withdrawing from the course after the halfway point. In year 4 this policy was modified to allow students to withdraw after the halfway point if they spoke with a school counselor first. These policy changes did not impact the course structure but rather the postinstruction class population.

In year 5 the primary textbook used in the course was changed, which (according to the instructor) adversely affected the timing of the lessons. Only one small-group problem-solving task was completed during the lecture sessions.

Year 6 brought a college-wide switch to semester-long courses. The general physics course consequently went from spanning three quarters to two semesters. This resulted in a course that was one week shorter than before. In year 7 the mathematics prerequisite courses were not enforced, and the small-group problem-solving tasks were eliminated completely. Several questions from the University of Massachusetts Physics Education Research Group were incorporated into the *Peer Instruction* sessions.

3. Comparing school 1 and school 2

Though we often compare similar forms of instruction at different institutions, such a comparison is suspect with the data presented in this study. Both institutions used many different forms of instruction, but school 2 classes approached mastery far more often than did school 1 classes. If one hypothesizes that using several different research-based instructional strategies is an appropriate way to improve student learning, the data are inconclusive: results from school 2 support this mindset, results from school 1 do not. One can imagine why the issue is confused: with so many different curricula in use in a single course, students might be having a hard time connecting the elements to each other.

We can speculate on some of the differences in instruction. It might be that the most important difference is the instructor. Having one instructor in lecture, problem-solving activities, and laboratory, like at school 2, means that students are getting a consistent message about what is important in the course and what leads to learning and success. In contrast, having a different instructor for lecture, tutorial, and laboratory activities, like at school 1, might lead to confusion for the students. For example, the different instructors might not highlight the connections between different instructional tools, and students might not arrive at a meaningful understanding of the common physics within the different curricula.

APPENDIX B: ADDITIONAL MODEL ANALYSIS PLOTS

Figures 8 and 9 provide additional model analysis plots of FMCE data from schools 1 and 2.

FIG. 8. Additional model analysis plots of FMCE data from school 1.

FIG. 9. Additional model analysis plots of FMCE data from school 2.

- [1] L. Bao, K. Hogg, and D. Zollman, Model analysis of fine structures of student models: An example with Newton's third law, *Am. J. Phys.* **70**, 766 (2002).
- [2] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010103 (2006).
- [3] 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).
- [4] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, *Phys. Rev. ST Phys. Educ. Res.* **4**, 020101 (2008).
- [5] 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).
- [6] K. C. Cummings, J. Marx, R. K. Thornton, and D. Kuhl, Evaluating innovation in studio physics, *Am. J. Phys.* **67**, S38 (1999).
- [7] E. F. Redish, Millikan Award lecture (1998): Building a science of teaching physics, *Am. J. Phys.* **67**, 562 (1999).
- [8] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010105 (2009).
- [9] E. F. Redish, J. M. Saul, and R. N. Steinberg, On the effectiveness of active-engagement microcomputer-based laboratories, *Am. J. Phys.* **65**, 45 (1997).
- [10] E. F. Redish and R. N. Steinberg, Teaching physics: Figuring out what works, *Phys. Today* **52**, No. 1, 24 (1999).
- [11] See Ref. [15] for a discussion of various methods for calculating the normalized gain of a class and their implications.
- [12] E. F. Redish, A theoretical framework for physics education research: Modeling student thinking, in *Proceedings of the International School of Physics “Enrico Fermi,” Course CLVI*, edited by E. F. Redish and M. Vicentini (IOS Press, Amsterdam, 2004), pp. 1–56.
- [13] T. Coladarci, C. D. Cobb, E. W. Minium, and R. C. Clarke, *Fundamentals of Statistical Reasoning in Education* (John Wiley & Sons, Inc., New York, 2004).
- [14] For our analyses we let $g = 1$ in these cases.
- [15] L. Bao, Theoretical comparison of average normalized gain calculations, *Am. J. Phys.* **74**, 917 (2006).
- [16] J. D. Marx and K. Cummings, Normalized change, *Am. J. Phys.* **75**, 87 (2007).
- [17] If the scores of each data set are reported as the number correct, then the leading term should be $1/\max$ to appropriately convert to a scale for normalized gain.
- [18] We find this to be the case with student understanding of Newton’s third law [4,20].
- [19] S. R. Sommer and R. S. Lindell, Accounting for errors in model analysis theory: A numerical approach, *AIP Conf. Proc.* **720**, 49 (2004).
- [20] T. I. Smith and M. C. Wittmann, Comparing three methods for teaching Newton’s third law, *Phys. Rev. ST Phys. Educ. Res.* **3**, 020105 (2007).
- [21] We continue using the “year X ” labeling system for years in which multiple sections were taught (rather than an arbitrary scheme) to allow interested readers to compare the data presented here with the instructional modifications described in Appendix A.
- [22] The Energy cluster questions were not administered in either of the sections during year 1.
- [23] Year 6 also appears to be similar to year 1b on both Newton III model plots, but for now we are mainly concerned with comparing these results to those in Fig. 6 and Table V.
- [24] Incidentally, Bao proposed a definition for a measure of normalized change between two data points on a model plot [25]. Initial tests with data from school 1 suggest that bar graphs of these values look very similar to those in Fig. 3, but a full presentation is beyond the scope of this paper.
- [25] L. Bao, Ph.D. dissertation, University of Maryland, 1999 (unpublished), available at <http://physics.umd.edu/perg/dissertations/>.
- [26] D. R. Sokoloff and R. K. Thornton, Using interactive lecture demonstrations to create an active learning environment, *Phys. Teach.* **35**, 340 (1997).
- [27] L. C. McDermott and P. S. Shaffer, and the Physics Education Group at the University of Washington, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).
- [28] M. C. Wittmann, R. N. Steinberg, and E. F. Redish, *Activity-Based Tutorials Volume 1: Introductory Physics* (John Wiley & Sons, Inc., New York, 2004).
- [29] A. Elby, R. E. Scherr, T. McCaskey, R. A. Hodges, E. F. Redish, D. Hammer, and T. J. Bing, Open Source Tutorials in Physics with Teaching Assistant Development (2007), <http://www.spu.edu/depts/physics/tcp/tadevelopment.asp>.
- [30] M. Wells, D. Hestenes, and G. Swackhammer, A modeling method for high school physics instruction, *Am. J. Phys.* **63**, 606 (1995).
- [31] G. Davenport, senior undergraduate thesis, University of Maine (unpublished).
- [32] This half of the student body is not correlated in any way with the half that comprised the experimental group during the modeling method study, nor do either of these groups correspond with either of the experimental groups for the Newton III study.
- [33] M. Dudley and E. Bianchi, The influence of studying equipotentials in mechanics when learning electrostatics, in *Proceedings of the 2005 Summer Meeting of the American Association of Physics Teachers, Salt Lake City*.
- [34] R. K. Thornton and D. R. Sokoloff, Learning motion concepts using real-time, microcomputer-based laboratory tools, *Am. J. Phys.* **58**, 858 (1990).
- [35] D. R. Sokoloff, R. K. Thornton, and P. W. Laws, *RealTime Physics* (John Wiley & Sons, Inc., New York, 1998).
- [36] E. Mazur, Peer Instruction: A User’s Manual, *Series in Educational Innovation* (Prentice-Hall, Upper Saddle River, NJ, 1997).
- [37] Research into gender biases in physics suggests that gender-neutral scenarios in examples and assignments may be beneficial to both male and female students [38].
- [38] L. McCullough, Gender, math, and the FCI, in *2002 Physics Education Research Conference*, edited by S. Franklin, K. C. Cummings, and J. Marx (PERC Publishing, New York, 2002).
- [39] Tycho Web-based Homework System, <http://research.physics.illinois.edu/per/Tycho.html>.