Beyond normalized gain: Improved comparison of physics educational outcomes

Elaine Christman, Paul Miller, and John Stewart[*]

West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506, USA

(Received 4 April 2023; accepted 25 March 2024; published 9 April 2024)

This study proposes methods of reporting results of physics conceptual evaluations that more fully characterize the range of outcomes experienced by students with differing levels of prior preparation, allowing for more meaningful comparison of the outcomes of educational interventions within and across institutions. Factors leading to variation in post-test scores on the Force and Motion Conceptual Evaluation (FMCE) across different instructors, semesters, and course models in a sample collected in introductory calculus-based mechanics at a large, eastern land-grant university were examined. The sample was collected over nine years and contains a total of $N = 4409$ matched pretest and post-test records. The data showed a systematic semester-by-semester variation in both pretest scores and ACT or SAT mathematics percentile scores. Neither the normalized gain nor Cohen's d removed the semester-to-semester variation observed in post-test scores. The local average curve plotting post-test scores against pretest scores, which we call a conceptual growth curve, allowed for the characterization of outcomes for students with different pretest scores. Regression models were used to produce an approximation to this curve. By using either the full curve or a mathematical approximation developed through linear regression, the post-test score that would be observed if a class enrolled students with a given level of prior preparation measured by pretest scores can be predicted. This predicted post-test score can then be used to calculate the predicted normalized gain if desired. These methods rely on using the natural variation of incoming student preparation at one institution to predict how a class would perform if it enrolled students with different prior preparation. The study presents an example of converting the outcomes at an institution with a weakly prepared student population to the outcomes which would have been observed if the course enrolled a more prepared student population; converting the outcomes for a different student population dramatically changed the interpretation of how the class studied was functioning.

DOI: 10.1103/PhysRevPhysEducRes.20.010123

I. INTRODUCTION

Assessing student learning is fundamental to improving the quality of instruction and in evaluating the performance of instructors. Comparing the performance of instructional models across classes and institutions is vital to identifying instructional models that promote learning and in arguing for the broad dissemination of those pedagogies. The incoming preparation of students both in general and in physics varies among classes at the same institution and may vary broadly between institutions making these comparisons difficult.

Within physics, standardized research-based conceptual instruments have become one of the most commonly used assessments of physics instruction. For an extensive review of research-based instruments in physics see Madsen *et al.* [1]. Popular instruments include the Force Concept Inventory (FCI) [2] and the Force and Motion Conceptual Evaluation (FMCE) [3]. These instruments measure a student's conceptual understanding of Newton's laws. The assessment is given early in the semester as a pretest to gauge student incoming preparation, then late in the semester as a post-test to measure student knowledge after instruction. Pretest and post-test experimental designs have been common in physics education research (PER) since its inception; many excellent summary articles provide an overview of the field and of subfields where their use is particularly widespread [4–6].

In 1985, Halloun and Hestenes used pretest and post-test data to show little conceptual understanding was gained in a traditional physics class [7]. This work ultimately led to the development of the FCI, the first broadly adopted PER research-based instrument. In an effort to determine the effectiveness of reformed instruction, Hake collected FCI pretest and post-test data from 62 courses across a broad variety of institutions [8]. To compare learning across these institutions with substantially different pretest scores, Hake

[*] jcstewart1@mail.wvu.edu

plotted the average gain (post-test to pretest) for each institution against the average pretest score. Examination of the resulting plots led Hake to propose the normalized gain, g, the ratio of the average gain to the average total possible gain (100%—pretest) as shown in Eq. (1), as a useful statistic to compare gains across diverse institutions:

$$g = \frac{\langle Post \rangle - \langle Pre \rangle}{100 - \langle Pre \rangle}, \tag{1}$$

where pretest is abbreviated as *Pre*, post-test as *Post*, and both pretest and post-test are scored out of 100%. The average of a variable X is represented by $\langle X \rangle$.

Hake stated [8] "I infer from features (figures in [8]) that a consistent analysis over diverse student populations with widely varying initial knowledge states, as gauged by $\langle pretest \rangle$, can be obtained by taking the normalized average gain $\langle g \rangle$ as a rough measure of the effectiveness of a course in promoting conceptual understanding. This inference is bolstered by the fact that the correlation of $\langle g \rangle$ with $\langle pretest \rangle$ for the 62 survey courses is a very low +0.02." The Hake study was central to the effort to encourage the adoption of interactive methods of physics instruction; the influential nature of the study also led to the broad adoption of the normalized gain in PER.

The use of research-based instruments has grown to the extent that large studies aggregating data from multiple institutions are now possible. Von Korff *et al.* [9] gathered the results of studies administering either the FCI or FMCE from 1995 to 2014 producing a sample containing 50 000 students. A synthesis of this data demonstrated that interactive instruction produced superior normalized gains when compared to traditional instruction. Freeman *et al.* showed that interactive instruction was superior to traditional instruction in producing learning gains and promoting student success; this result held across a variety of science, technology, engineering, and mathematics (STEM) domains [10].

Since its introduction, substantial evidence has accumulated that the normalized gain does not completely correct for differing student prior preparation. Coletta and Phillips showed that normalized gain scores were correlated with FCI pretest scores (correlation coefficient $r = 0.33$) [11]. This result has been replicated by a number of other studies [12,13]. The current study shows this correlation between normalized gain and pretest scores is also found for the FMCE. Coletta *et al.* went on to establish that some of this relation could be explained by a correlation between standardized test scores (SAT scores) [14] and the normalized gain. Several recent studies have reported correlations between either FCI or FMCE pretest and standardized test scores (either the ACT or SAT) [15–17].

While broadly reported in PER for many years, the normalized gain has recently become somewhat controversial. The statistic is generally only reported in PER, and as such, if it is not accompanied by more broadly used measures such as pretest scores and post-test scores, its use may make PER studies difficult to interpret by the broader education research community. The statistic has been inconsistently reported [12] with some studies first calculating the averages $\langle Pre \rangle$ and $\langle Post \rangle$ then using Eq. (1) to calculate the normalized gain while other studies have first calculated a normalized gain for each student, g_i, [Eq. (2)], then averaged this result to produce the average normalized gain, $\langle g_i \rangle$,

$$g_i = \frac{Post_i - Pre_i}{100 - Pre_i}, \tag{2}$$

where Pre_i is the pretest score of student i and $Post_i$ the post-test score of student i; both pretest and post-test are scored out of 100%. The two methods do not yield equivalent results [18]. In the current work, we use the method applied by Hake in Eq. (1). If the student-level method is used, singularities may occur which led Marx and Cummings to propose an alternate statistic, the normalized change [13].

In an analysis of conceptual inventory data from biology, chemistry, and physics, Nissen *et al.* found that normalized gain was positively biased in favor of populations with higher pretest scores, which they suggested resulted in an overestimation of course-level gender bias. They proposed discontinuing the use of the normalized gain in favor of Cohen's d between the pretest and the post-test

$$d = \frac{\langle Post \rangle - \langle Pre \rangle}{s_P}, \tag{3}$$

where s_P is the pooled standard deviation of the pretest and post-test. The pooled standard deviation is the sample size weighted square average of the pretest standard deviation, s_{pre}, and the post-test standard deviation, s_{post}, as shown in Eq. (4)

$$s_P = \sqrt{\frac{(n_{pre} - 1)s_{pre}^2 + (n_{post} - 1)s_{post}^2}{(n_{pre} + n_{post} - 2)}}, \tag{4}$$

where n_i is the sample size. They argued that using Cohen's d mitigates both ceiling and floor effects [12].

This suggestion was strongly opposed by Coletta and Steinert, who argued that normalized gain is not prescore biased. They suggested retaining normalized gain as a measure of the effectiveness of pedagogical approaches while proposing that scores on the Lawson Classroom Test of Scientific Reasoning, ACT, or SAT must also be considered when making comparisons [19].

The primary innovation of the present work is the use of the natural variation of student prior preparation and outcomes at the institution studied to allow the prediction of how classes would perform if they enrolled a different student population. This work proposes reporting the functional relation between post-test scores and pretest scores to make use of this natural variation. This functional relation may either be reported using a visualization of the relation of pretest to post-test or by reporting the mathematical function relating pretest to post-test.

Reporting such measures of preparation or ability for student populations studied is not universal in published PER studies. The likelihood of authors noting the impact sample characteristics have on the generalizability of published results has increased as the field has matured [20], and while tools such as PhysPort allow instructors to construct histograms to compare their class outcomes on conceptual inventories to those of similar classes [21], these do not facilitate comparison across classes with differing incoming preparation. While visualization of pretest and post-test score distributions is common in published studies, the visualization of the relation of pretest scores to post-test scores is not. Visualizations of data allow researchers to leverage both computation and human cognition to make sense of large, heterogeneous datasets [22].

The most common graphical representation of pretest or post-test scores that moves beyond reporting of overall averages presents histograms of pretest and post-test scores [23–26]. Multiple studies [27,28] have used a histogram to provide a more nuanced characterization of gender differences on the FMCE by showing average postscore for a range (bin) of pretest scores disaggregated by gender. This representation is related to that proposed in the current work. Unfortunately, the nonlinear binning used does not allow the extraction of the full pretest or post-test response curve. Histograms have also been used to represent the distribution of items scores [29].

While graphical representations of the relation of pretest and post-test scores are rare in PER, graphical representations of the relation of pretest scores to individual items responses in the form of item characteristic curves for nominal item response theory [30] or item response curves [31,32] have been reported. Stacked histograms have also been used to represent pretest to post-test changes to item responses [3].

A representation related to the heat map used in Fig. 3 was used by Thornton *et al.* [33] to show the relation of FMCE scores to FCI scores.

Reporting the mathematical relation of pretest to post-test as a linear regression is fairly common in PER [15–17,34]. This is generally done in studies investigating the effect of a set in independent variables including pretest scores on post-test scores. The regression is reported to characterize the relation of the variables, not as a means to report the variation in the data. As such, the regression equations reported generally contain additional variables beyond pretest scores making them difficult to use to predict student outcomes from data drawn from different institutions. The studies rarely report the additional quantities needed to compute confidence intervals for predicted post-test scores (see Supplemental Material [35]). Further, these studies seek to optimize model fit, selecting models that fit well in regions with many pretest observations, but for reporting purposes, the models should fit well over the range of pretest scores. In the present work, we will need to include a quadratic term in pretest to fit the data well over its full range; higher order powers of pretest scores are rarely explored.

A. Research questions

The purpose of this study is to propose methods that allow student conceptual learning to be compared across classes at the same institution and to allow published research results to be evaluated as to determine if they are likely to increase student success when implemented at a different institution.

This study seeks to answer the following research questions.

RQ1 Do either the normalized gain or Cohen's d allow productive comparison between classes at one institution if the student characteristics vary between classes?

RQ2 How can the natural variation of student prior preparation within a course be used to compare classes enrolling students with differing levels of prior preparation?

RQ3 How can the natural variation of student outcomes within a course be used to evaluate the efficacy of published PER results when transferred to a local context?

II. METHODS

A. Sample

This study was performed at a land-grant university in the eastern United States with total undergraduate enrollment of 20 500 in fall 2020. Data were collected in the introductory calculus-based mechanics class taken by scientists and engineers from Fall 2011 to Fall 2019. The general demographic composition of the university in 2019 was 82% White, 4% Black or African American, 4% Hispanic/Latino, 4% nonresident alien, 4% two or more races, with other groups 2% or less. The 25th percentile to the 75th percentile range of ACT composite scores range was 21 to 27 [36]. This is equivalent to a range of composite scores of 59th to 85th percentile. Pell grants are available to students of lower socioeconomic status; 31% percent of undergraduate students were Pell eligible. This study collected FMCE pretest and post-test scores; students received assignment grades in the course for good faith efforts on these assessments. Standardized test scores were accessed from institutional records.

B. The FMCE

The FMCE [3] is a 43-item multiple-choice instrument (excluding the four energy questions added after its initial publication). The instrument measures a student's understanding of Newton's laws and one-dimensional kinematics. Thornton *et al.* [33] proposed a modified scoring methodology, which eliminated scoring of some questions typically answered correctly even by students with non-Newtonian beliefs and one question not consistently answered correctly by experts. Certain related questions are scored as groups, producing a total instrument score of 33. The present study uses the modified scoring method.

C. The instructional environment

The class studied was presented in two instructional models over the period studied. Data were collected for 18 semesters numbered 1 to 18. The first model, the learning assistant (LA) instructional model, was in place from semester 2 to semester 9. This model presented four 50-min lectures each week. The lectures were taught by a variety of faculty, many of whom used some interactive engagement method. Students also enrolled in a two-hour required laboratory section. The lab section was split into two hour-long halves. The first half was led by undergraduate learning assistants [37] who helped the students work through a lesson from the University of Washington's *Tutorials in Introductory Physics* [38]. The second half of the lab involved a traditional experiment overseen by graduate teaching assistants (TA). The LAs were required to enroll in a physics teaching methods course in which they received training in science teaching methods in general and specific instruction in presenting the upcoming week's lesson. The program was supported as part of a larger general science grant which allowed the LAs to be compensated for their efforts. Coordination between the lecture and laboratory parts of the course was variable and depended on the lecture instructor. This instructional model had to be abandoned when funding for the LA stipend was discontinued with the end of the grant.

From semester 10 to 18, an alternate model not requiring LAs to staff each lab was implemented. This model focused on coordinated lecture and laboratory instruction and is called the coordinated learning (CL) model in this work. This model shifted to three 50-min lectures and one 3-h lab each week. A single lead instructor assumed the course coordinator role and ensured that all lecture sections progressed on a fixed schedule in concert with laboratory sections. Lecture instructors were encouraged to implement Peer Instruction [39], and all lecture sections included clicker questions. Labs were modified to feature a combination of activities including whiteboarding of conceptual problems, TA-led demonstrations, group problem solving, hands-on inquiry activities, and traditional experiments. The LA program was modified to one requiring the election of the physics teaching methods class for credit; however, LAs were no longer compensated, and LAs could not be provided for each laboratory section. The *Tutorials in Introductory Physics* were abandoned because of their cost to students and because they could not be modified to fit the instructional setting. The materials were replaced with a combination of some modified elements of the *Open Source Tutorials in Physics Sensemaking* [40] along with other interactive activities to produce a coherent 3-h session.

Semester 1 was a control semester for the LA program and is not reported because only one semester of data was collected. It is not included in the graphs or other analyses.

D. Variables

The FMCE was given as a pretest early in the semester and as a post-test during the last week of the semester. The students received credit for a good faith effort. Standardized test scores were accessed from institutional records. Students reported either ACT or SAT scores; the mathematics subscore was converted to a percentile using conversions published by the testing companies. If both scores were reported, the percentiles were averaged. The quantity will be represented by the variable ACTM%.

III. RESULTS

Table I presents descriptive statistics for the class studied. The quantities are reported as mean ± standard deviation. Both the normalized gain and Cohen's d are calculated at the class level and as such no standard deviation could be calculated. The statistics are reported for the aggregated overall sample and disaggregated by instructional model.

For semester 1, the control semester using a traditional instructional model, 175 students completed the pretest and post-test. The average pretest score was 23.6% ± 19%, the average post-test score as 41.1% ± 27%, and the average ACTM% percentile score was the 80.1 ± 15. This yielded a normalized gain of 22.9%.

A. Variation of course outcomes

Figure 1 plots the various measures of prior preparation and class achievement by semester. Semesters are

TABLE I. Descriptive statistics. The mean ± standard deviation is presented.

Variable	Overall	LA model	CL model
ACTM %	79 ± 15	77 ± 16	80 ± 15
Pretest %	23 ± 18	22 ± 17	24 ± 19
Post-test %	48 ± 28	49 ± 29	47 ± 28
Normalized gain %	36 ± 32	38 ± 32	34 ± 32
Cohen's d	1.06	1.16	0.98
N	4409	2088	2321

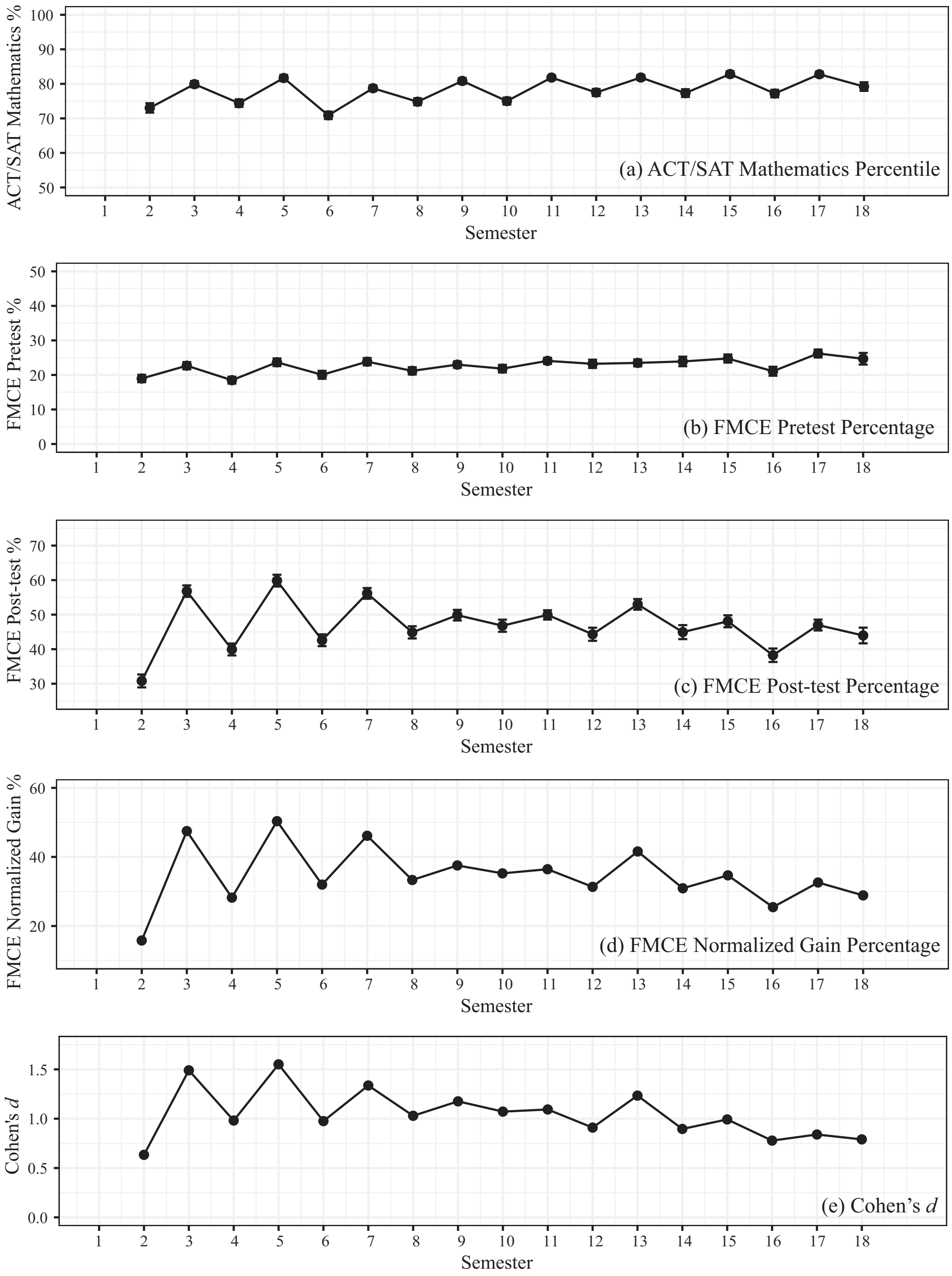

FIG. 1. The average ACT/SAT mathematics percentile (ACTM%), FMCE pretest, post-test, normalized gain, and Cohen's *d* per semester.

numbered from 1 (Spring 2011) to 18 (Fall 2019). Spring semesters are odd numbers; fall semesters even numbers. Error bars are provided for most plots; these represent the standard error of the mean. Normalized gain is calculated at the semester level using Eq. (1). Cohen's *d* between the pretest and post-test and the normalized gain require semester-level variables, and as such, no error bars could be calculated. For fair comparison while maintaining ease of reading, Figs. 1(a) to 1(d) are plotted in a 50% wide range.

All plots show an oscillation from spring to fall, with spring (odd) semesters having both higher incoming ACT or SAT mathematics percentile (ACTM%) scores and FMCE pretest scores and higher outcome values on the FMCE post-test, FMCE normalized gain, and Cohen's d. This is likely a result of the prerequisite structure of the course studied. The course has Calculus 1 as a prerequisite; students ready to enroll in Calculus 1 in their fall freshman semester enroll in the course studied in their spring freshman semester. This is the university's intended course progression for most students in the course as outlined in published four-year degree plans. However, many students taking the class in fall semesters were not eligible to take Calculus 1 upon entering college; these students were not "math ready" and thus delayed enrollment in introductory physics.

The pretest and ACTM% plots clearly show that instructors teaching in the spring semester have a more academically prepared student population; these instructors produce superior academic outcomes measured by FMCE post-test scores. Because of the differences in incoming student preparation, it is difficult to determine if differences in academic outcomes result from differences in instruction or differences in student population. It is thus challenging to use post-test scores fairly to evaluate instruction. It is also very difficult to evaluate the results of educational reform against this background of student variation.

There are not instructional reasons for this variation by semester to exist. The class is presented in the same format in both the spring and the fall semesters. Instructors are also fairly randomly assigned among semesters.

Figure 1 also clearly shows that neither the normalized gain nor Cohen's d does much to remove this semester to semester variation. If either the normalized gain or Cohen's d were effective at controlling for varying student characteristics, one of their plots would be randomly distributed around some value; they are not. As such, neither metric is useful in providing an assessment of instruction that is independent of variation in student characteristics. Because Cohen's d seems the have the same flaws as the normalized gain which is more widely reported in PER, we will focus on normalized gain for the remainder of this study.

Figure 2 plots the normalized gain against both FMCE pretest score and ACTM%. Each point represents a single lecture section. A regression line has been added to each plot. This line was calculated using the student-level data aggregating all lecture sections and all semesters. The regression lines show that, due to the variation in the student population, one could expect a variation of normalized gain of 15 percentage points between the lowest and highest preforming classes.

The level of relation between pretest, normalized gain, and ACTM% can be characterized by the correlation coefficient r. The pretest and ACTM% ($r = 0.30$), the post-test and ACTM% ($r = 0.43$), and the normalized gain and ACTM% ($r = 0.39$) were all significantly correlated ($p < 0.001$). All were medium effects by Cohen's criteria [41]. The correlation between these variables is well established in the literature with many studies showing both ACTM% and pretest scores are important in models predicting post-test scores [15,17]. These levels of correlation between the normalized gain and ACTM% were similar to but somewhat smaller than those observed by Coletta *et al.* [14] between SAT scores and FCI normalized gains in both university and high school students. These results suggest the correlation of standardized test scores to conceptual inventory results is fairly general. The student-level normalized gain was also significantly correlated with pretest scores ($r = 0.43$, $p < 0.001$) which contradicts Hake's observation that they were not correlated [8].

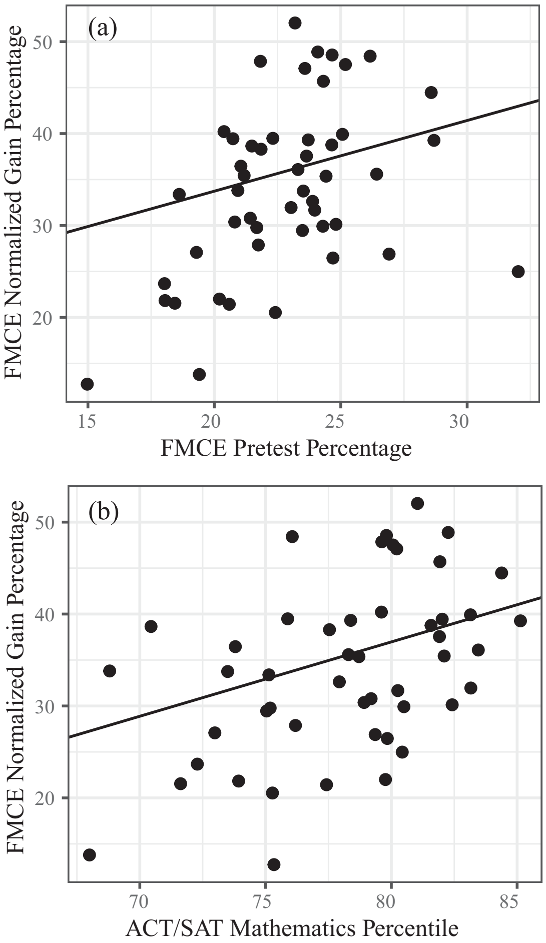

FIG. 2. (a) The normalized gain plotted against FMCE pretest percentage. (b) The normalized gain plotted against ACT/SAT mathematics percentile (ACTM%). Each point represents the average of an individual course lecture section. The line is the regression line using the full dataset.

B. Representing the variation in student outcomes

The course studied has been presented in many individual lecture sections taught by a variety of instructors

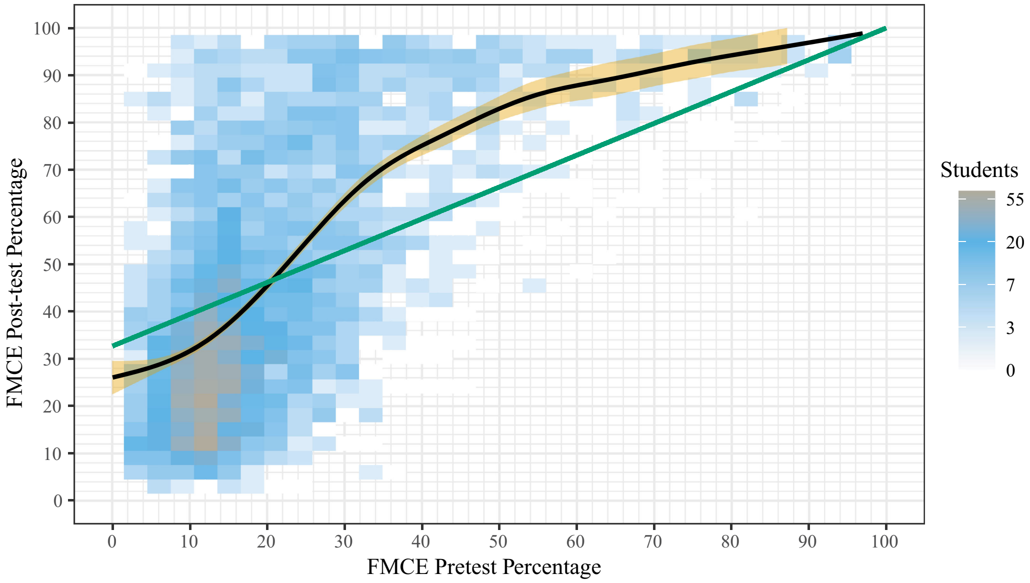

FIG. 3. FMCE post-test scores vs pretest scores. The local average is plotted in black with its 95% confidence interval. The Hake Model for the normalized gain is plotted in green. The background is a heat map showing the density of the data at each point.

over the period studied. The variation in student preparation between sections makes it very difficult to fairly compare the outcomes of two instructors or two pedagogical models. We need to be able to answer questions such as “What are the predicted post-test outcomes of class A given its student composition based on the overall performance of the course?” While one cannot change the student composition of class A, the course as a whole over the time studied has enrolled students with a broad range of prior preparation. The range of outcomes of these students can be used to predict the expected outcome of class A given the prior preparation of its students. To accomplish this, a representation of how post-test scores change with pretest scores (and other variables) is needed.

A prior study of this course using only the CL instructional model, but with access to a substantial number of control variables examined the relation of post-test scores to pretest scores [17]. Linear regression analysis showed that the most important factors in predicting post-test score were pretest score explaining 44% of the variance in post-test score followed by ACT/SAT scores explaining an additional 6% of the variance. The remaining 32 variables together explained only 4% additional variance. As such, it seems reasonable to start with a representation of how post-test scores vary with pretest scores. Linear regression analysis of this data is explored in Sec. III C.

The natural variation of student incoming preparation can be used to build a model predicting post-test scores from pretest scores and other variables. As such, the variation in student preparation in the course overall is used to predict how a class section should have performed given the prior preparation of its student population. One method to build such a model that does not rely on any assumptions about the statistical properties of the distribution of the variables of interest is to report a plot of post-test scores against pretest scores including a local average curve. A local average curve computes the average of post-test scores for a narrow range of pretest scores possibly adding some smoothing. Many methods exist to produce a local average curve; some will be discussed in this work. Figure 3 shows a plot of post-test scores against pretest scores. The black curve in the figure represents the local average. We propose these curves be called “conceptual growth curves” (CGC). The underlying scatterplot is not shown as the large number of stacked data points results in a figure that is difficult to interpret. Instead a grid heatmap in the background of the plot depicts the numerosity of students in the dataset with each pretest or post-test score combination. As might be expected, there is a strong density of students earning scores near the average pretest or post-test score; however, there is also a substantial population of students away from the average. The Supplemental Material [35] contains a table showing the average post-test score for each of the 34 (0 to 33) possible pretest scores as well as the number of students with that pretest score. This table may be easier to interpret than the heat map.

There are many ways to calculate the local average curve. The most straightforward is the calculate the average post-test score for each pretest score. One can also compute the average of the post-test for a small range of pretest scores centered on each pretest score (this is sometimes called a moving average). Both methods are shown in the Supplemental Material [35]. Using an average including three consecutive pretest scores generated an excellent approximation to the curve shown in Fig. 3. In general, more sophisticated methods yield smoother curves. The local average curve in Fig. 3 was calculated with the geom_smooth function in the ggplot2 package in "R." For the sample size in this study, this package applies a general adaptive model to fit a set of splines to the data. For sample code to draw the curve, see the Supplemental Material [35]. Note, using the default settings of the "geom_smooth" algorithm in R will select a different fitting algorithm if $N < 1000$; the code in the Supplemental Material [35] forces the use of the general adaptive model for consistency.

Figure 3 can also help to explain why the normalized gain was not productive at the institution studied. The normalized gain hypothesizes a specific relation between pretest scores and post-test scores. For the normalized gain to be constant for different pretest scores, the CGC growth curve must be well fit by the model in Eq. (5).

$$\langle Post \rangle = 100 \cdot g + (1 - g) \cdot \langle Pre \rangle. \tag{5}$$

This results from solving Eq. (1) for $\langle Post \rangle$. The equation has been plotted in Fig. 3 as the "Hake model." It is a fairly poor approximation to the CGC. For the normalized gain to be used for comparison, one first needs to confirm that the Hake model is a good approximation to the CGC. We know of no instance, where the normalized gain was reported, that this crucial step was performed.

We note that the Hake model is an exceptionally reasonable model, perhaps the only reasonable model, if the only information available is the pretest and post-test averages. The Hake model interpolates between the point representing the average pretest and post-test score and the point (100, 100). We further note that, in general, the Hake model and the CGC will cross near the average of the pretest and post-test and the point (100,100); the Hake model is a chord of the CGC between these two points. If a higher pretest score had been observed, the chord becomes a better approximation for pretest scores greater than the average score. In the figure, if the pretest score had been 50%, the Hake model would be a very good approximation for scores greater than 50%, but an extremely poor model for scores less than this value.

To understand the use of a CGC to predict what outcome was expected from a class section, we examine the lecture sections with the lowest and highest pretest scores. The lowest pretest score class section, class L, had a pretest average of 15% for a post-test average of 25%. The highest pretest section, class H, had a pretest average of 32% for a post-test average of 48%. The CGC can be used to determine what post-test score the course produces on average for this pretest score. Reading the value of the CGC at 15%, for class L, a post-test score of 37% is predicted; class A under performed its expected average by 12%. For class H, a post-test score of 67% was predicted; class H also substantially under performed its predicted average by 19%. The same correction can be performed with the Hake Model yielding a predicted post-test score for class L of 43% and for class H of 54%. As might be expected from Fig. 3, the Hake model over predicts for very low pretest scores and under predicts very high pretest scores. As one can see, the error is substantial. From this, we can see that both class L and H perform more weakly than anticipated correcting for differences in pretest scores and that using the normalized gain produces substantial errors in the predicted results.

The figures presented in Sec. III A and prior work [17] suggest post-test scores may vary with both ACTM% and pretest scores. It is possible that pretest scores may fully explain the effect of variation of ACTM%. This possibility can be investigated by further plotting a CGC for each quantile of ACTM% scores; this analysis is shown in the Supplemental Material [35]. This analysis suggests ACTM% scores are important in addition to pretest scores in explaining post-test scores consistent with prior work [17].

Figure 4 illustrates how the CGC can be used to compare the efficacy of an educational intervention for the range of incoming student preparation to provide a more nuanced picture than that provided by a single general metric. A CGC for each instructional model is plotted, with LA in black and CL in orange. The shaded area represents the 95% confidence interval for each curve. The LA instructional model outperforms CL for all but students with extremely low pretest scores; however, the difference in model outcomes for the best-prepared students is minimal, with results of the LA model within the 95% confidence interval of the CL model. The overall effect will be quantified in the next section as mathematical models are built.

Beyond productively comparing classes at the same institution, the instructional personnel also need to make informed judgments about whether published educational reforms would be effective at improving local instruction. As such, they wish to answer the question, "How would the educational model presented in the published work perform if it enrolled students at the local institution?" Unfortunately, with the normalized gain failing to correct for student differences, this generally cannot be answered using existing reporting of PER results. The CGC allows the answering of a somewhat less useful but valuable question, "How would the local class perform if it enrolled a student population with similar characteristics to those in a published study?"

As an example of productively answering this kind of question, we use the course results of a recently published study. Salehi *et al.* [15] presented pretest and normalized gain results as well as standardized test scores for three institutions in an effort to understand the effects of prior preparation on final exam scores for a number of demographic groups. The largest sample reported ($N > 4000$), referred to as PM in the study, reported FMCE scores and

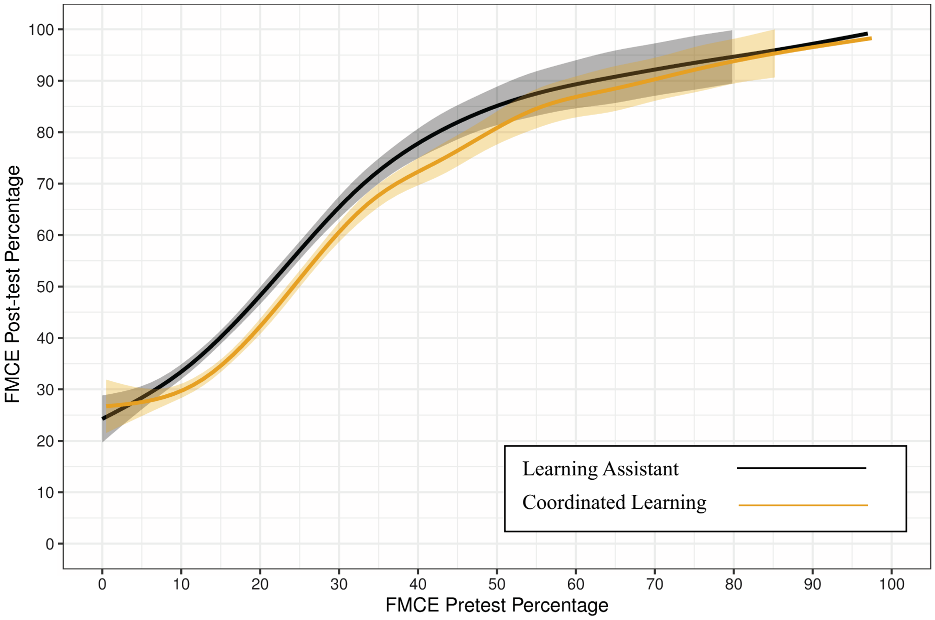

FIG. 4. Conceptual growth curve comparing LA instruction (black) with coordinated learning CL instruction (orange) including the 95% confidence interval.

ACT or SAT mathematics percentile scores. As in this study, a mixture of ACT and SAT results were available. Students at PM had an average ACT or SAT mathematics percentile score of 89% and pretest scores for multiple semesters ranging from 38% to 49%. The center of this range of pretest scores is used for comparison, $Pre = 43.5\%$. Post-test scores were not reported; however, PM achieved a normalized gain of 49% to 54%, the center of the range of normalized gain is 51.5%. Using the reported average normalized gain and pretest score, one can calculate that the post-test scores at PM were approximately $Post = 72.6\%$. This study was chosen as an example because enough parameters were reported for evaluation and because the FMCE was used; this study does not present the pedagogy used. As such, if analysis indicated that PM was performing substantially better than the local course, additional investigation would be needed to identify its instructional model.

Using the CGC in Fig. 3, one can determine that if the local course (the course studied) enrolled students with an average pretest score of 43.5%, then the CGC predicts the local class would produce an average post-test score of 76% yielding a normalized gain of 57%. As such, the local class would outperform the class in Salehi *et al.* [15] giving evidence that it would not be productive to try to determine the pedagogy used in this study so as to adopt the pedagogy to the local context. We note that this does not mean the class Salehi *et al.* would not produce superior results if it enrolled students like those in the local course; it is simply not possible to determine this with the data published. If the Hake model were used to perform the correction, a post-test score of 62% would be predicted, much less than that predicted by the CGC. This would inaccurately suggest that PM would outperforming the local model indicating that investigating adopting the pedagogy in the published study would be efficacious.

C. Modeling the conceptual growth curve

The CGC in Fig. 3 allows the prediction of post-test scores from pretest scores; however, its use requires the reading of the graph for each class of interest. This can become onerous with a large number of class sections and makes it hard to evaluate how all the factors work together. It becomes even more problematic when ACTM% scores or other variables are added. It would facilitate the use of the CGC to have a mathematical model of the curve. Beyond practical convenience, a mathematical model allows one to quantify the amount of variance explained by the model as well as the relative importance of the variables the model.

Hierarchical linear regression (HLR) was used to build a model of the CGC. HLR calculates a set of nested models where more complex models add variables to less complex models. This nested set of regression models is presented in Table II. All models were a significant improvement upon the model in which they were nested using a likelihood ratio test ($p < 0.001$). The curves resulting from models only involving pretest scores are shown in Fig. 5. The overall fit of the nested models is characterized by the Akaike information criterion (AIC) as shown in Eq. (6). AIC measures the relative information lost between the model and a "true" model while correcting for overfitting [42,43]. Smaller values of AIC represent better fitting models.

TABLE II. Post-test regression models. B is the regression coefficient, SE the standard error, β the standardized regression coefficient, t the t statistic, and p the probability a value as large or larger than t occurred by chance. The 95% confidence interval (CI) is also presented. All models are statistically significant improvements over the null model containing only an intercept ($p < 0.001$). Each nested model is a significant improvement over the model in which it is nested ($p < 0.001$).

	B	SE	95% CI	β	t	R^2	AIC
			Linear model				
(Intercept)	24.8687	0.52	[23.83, 25.91]	0.00	47.62	0.42	27 053
Pretest	1.0150	0.02	[0.98, 1.06]	0.65	56.63		
			Quadratic model				
(Intercept)	14.4190	0.83	[12.76, 16.08]	0.00	17.31	0.45	26 811
Pretest	1.8567	0.06	[1.74, 1.98]	1.19	33.18		
Pretest^2	−0.0104	0.00	[−0.0117, −0.0091]	−0.57	−15.83		
			Quadratic model with ACTM%				
(Intercept)	−16.5557	1.63	[−19.82, −13.30]	0.00	−10.16	0.51	26 363
Pretest	1.6254	0.05	[1.53, 1.73]	1.04	29.97		
Pretest^2	−0.0089	0.00	[−0.0077, −0.0102]	−0.49	−14.29		
ACTM%	0.4453	0.02	[0.41, 0.49]	0.24	21.74		
			Instructional model				
(Intercept)	−15.0754	1.63	[−18.34, −11.82]	0.09	−9.27	0.51	26 293
Pretest	1.6353	0.05	[1.54, 1.74]	1.05	30.39		
Pretest^2	−0.0090	0.00	[−0.0102, −0.0078]	−0.49	−14.51		
ACTM%	0.4584	0.02	[0.42, 0.48]	0.25	22.50		
Course model	−5.0945	0.60	[−6.29, −3.89]	−0.18	−8.53		

$$\mathrm{AIC} = 2k - 2\ln(L), \tag{6}$$

where k is the number of parameters and L is the likelihood function.

The simplest model predicting post-test score from pretest score is shown as the linear model in Table II which fits the equation $Post = B_0 + B_1 \cdot Pre$, where B_0 is the intercept and B_1 is the slope. This yields $Post = 24.9 + 1.02 \cdot Pre$; this line is drawn as the blue linear model in Fig. 5. This model explains 42% ($R^2 = 0.42$) of the variance in post-test score. The line does not qualitatively capture the shape of the CGC; the CGC is visually not a line. The fit to the local average can be improved by the addition of a term quadratic in pretest score to the regression model. This regression is shown as the quadratic model in Table II. The model fit was $Post = B_0 + B_1 \cdot Pre + B_2 \cdot Pre^2$ which yielded $Post = 14.4 + 1.86 \cdot Pre - 0.0104 \cdot Pre^2$. More digits were reported because of the range of the Pre^2 term. This model explained 45% of the variance in post-test score and is shown as the orange curve in Fig. 3. This model provides a good approximation to the local average curve except at very low pretest scores. A model adding a cubic term did not improve the visual fit of the model.

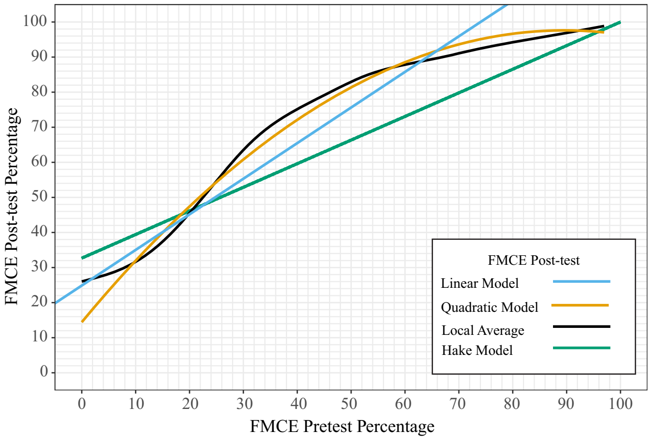

FIG. 5. FMCE post-test scores vs pretest scores. The local average is plotted in black. The Hake model for the normalized gain is plotted in green. The linear model is plotted in blue. The quadratic model is plotted in orange.

This model can be improved somewhat by including standardized test scores. These scores are available for most students at many institutions and are often reported in PER studies. The quadratic model using both pretest score and ACTM% score is shown as quadratic model with ACTM% in Table II. This model explained 51% of the variance in post-test score, $R^2 = 0.51$. The model fit was $Post = B_0 + B_1 \cdot Pre + B_2 \cdot Pre^2 + B_3 \cdot \mathrm{ACTM\%}$ which yielded $Post = -16.6 + 1.63 \cdot Pre - 0.0089 \cdot Pre^2 + 0.45 \cdot \mathrm{ACTM\%}$.

Linear regression can also quantify the general difference of the two instructional models. Adding the dichotomous variable course model (0 = LA, 1 = CL) to the quadratic model with ACTM% is shown as the instructional model in

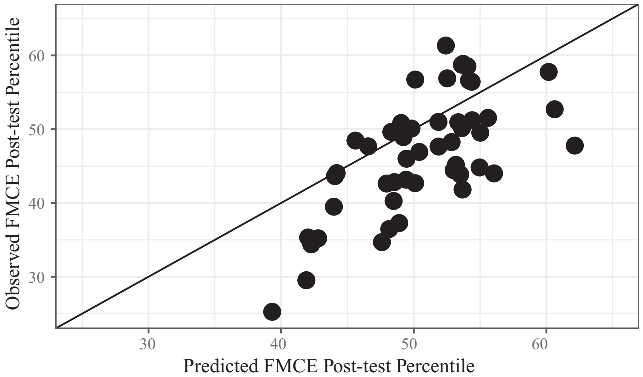

FIG. 6. Predicted FMCE post-test percentage vs observed FMCE post-test percentage by class section. The line has slope one. Classes above the line outperformed predictions; classes below the line under performed predictions.

Table II. The regression coefficient of the course model variable measures the difference between the two models; the LA model produced on average post-test scores which were 5.1% higher than the CL model.

Using the models presented in Table II allows the department to partially correct for the variation in student characteristics by class section. This allows the department to develop a more accurate picture of the instructional success of different faculty (allowing help to be directed where needed) and to more accurately characterize future course reforms. Figure 6 plots this predicted post-test score against the observed post-test score for each class section in the period studied. A line of slope one has been added to the figure. Class sections above this line outperform predictions. Many classes are near the line and perform approximately as expected. Some are well above the line; the methods of these instructor may be worth emulation. Some are substantially below the line indicating some additional support is needed.

The quadratic model with ACTM% model in Table II shows that general academic factors such as ACT and SAT scores are important to predicting student conceptual learning beyond pretest scores; however, the additional variance explained by these factors is small compared to that explained by pretest score alone. It also shows that neither pretest scores nor ACTM% scores fully account for the variation in post-test scores at this institution.

Using the regression models in Table II, the average post-test score at the institution studied can be modified to reflect the post-test score which would have been expected if the prior preparation of the class studied mirrored that of another institution, allowing for comparison with results at other institutions. The core insight of this method is that while two classes at different institutions may enroll wildly different students on average, it is likely that the less selective institution enrolls some students like those of the more selective institution. For example, institution PM in the study by Salehi *et al.* [15] had an ACTM% of 89% and average pretest score of $Pre = 43.5\%$. The regression model using only pretest score (quadratic model Table II) predicts a post-test score of $Post_{\mathrm{pred}} = 14.4 + 1.86 \cdot 43.5 - 0.0104 \cdot 43.5^2 = 75.6\%$; once the predicted post-test score is calculated, then a predicted normalized gain could be calculated as $g = 100\% \cdot (75.6\% - 43.5\%)/(100\% - 43.5\%) = 57\%$. If ACTM% scores are used as well, the regression model (Quadratic Model with ACTM% Table II) predicts a post-test score of $Post_{\mathrm{pred}} = -16.6 + 1.63 \cdot 43.5 - 0.0089 \cdot 43.5^2 + 0.46 \cdot 89 = 78.4\%$ would have been observed if the students from the Salehi study were enrolled in the class studied. The predicted normalized gain is then $g = 100\% \cdot (78.4\% - 43.5\%)/(100\% - 43.5\%) = 62\%$. This again exceeded the normalized gain reported for PM of 51.5% and commensurate with the 61% normalized gain calculated in the previous section using the CGC. Again, this suggests locally adopting the pedagogy implemented at PM would not improve local instruction.

The instructional model variable produced a statistically significantly better model in Table II, but explained little additional variance. The Supplemental Material [35] presents further exploration of this and other variables local to the class studied. In general, the additional variables explain little additional student-level variance, but did improve the semester-level fits.

Care should be used when reporting a regression or other mathematical model to represent the CGC. If using a local average or binning, the CGC presents an average or representative value for each possible pretest score. This allows a CGC to represent the data (with error) even in regions where relatively little data are available. The CGC then represents a fairly high dimensional model with $k + 1$ parameters where k is the number of items. For the FMCE using the scoring method suggested by Thornton *et al.*, $k = 33$; as such, the CGC fits 34 parameters. The models of this section are far more parsimonious with the most complex model (instructional model) estimating only 7 parameters. This reduction in the parameters fit comes with a cost; the Quadratic Model (Table II) underestimates the CGC by approximately 5% at moderate pretest scores and overestimates it by 10% at very low pretest score. If ordinary least squares is used to fit the model, model fit around the average pretest will be prioritized. This may cause the model to fit poorly away from the average. For use in correcting post-test results for student differences, the model should fit over the entire range of the data. As such, the higher dimensional CGC is preferable. The mathematical model of the CGC should be checked to ensure it fits well over a wide range of the data.

IV. DISCUSSION

This study investigated three research questions. Many of the results have been discussed in previous sections; the most important will be summarized.

RQ1: Do either the normalized gain or Cohen's d allow productive comparison between classes at one institution if the student characteristics vary between classes? This study added to a substantial body of evidence indicating that the normalized gain does not sufficiently correct for differences in student prior preparation. Coletta *et al.* [14] showed that FCI normalized gain scores were related to ACT or SAT scores 15 years ago suggesting that the normalized gain did not fully correct prior preparation. This work was supported by several additional studies [12,13]. Multiple recent studies have established relations between pretest scores and standardized test scores. This work added additional support for the correlation between both pretest scores and ACTM% and normalized gain as shown in Fig. 2. Normalized gain was significantly ($p < 0.001$) correlated with ACTM% $r = 0.39$ and with pretest score $r = 0.43$. These correlations were somewhat smaller but in the same range as those reported by Coletta and Phillips [11].

The current work adds to the evidence that normalized gain does not fully account for student differences while providing some further nuance as to why this is the case. The failure to eliminate the spring to fall variation in Fig. 1 provides visual evidence of the failure of the normalized gain to eliminate the effects of student variation. The introduction of the CGC allowed further understanding of the approximation involved in using normalized gain. With the CGC, one realizes that Hake proposed a specific model of the relation between pretest and post-test scores, called the Hake model above [as shown in Eq. (5)]. This model is a reasonable choice if only limited information about the student response is available; that is if one only knows the pretest and post-test average not the distribution of scores. This realization allows one to graphically determine if the Hake model is a good fit to the actual response; both the Hake model and the actual response, the CGC, are plotted in Fig. 3. For the institution studied, there is a substantial difference between the Hake model and the CGC over a broad range of pretest scores. Comparison of Hake-model-corrected and CGC-corrected post-test scores in Sec. III B gives additional evidence that the normalized gain does not accomplish what it was introduced to do, allowing the comparison of post-test scores for classes with different levels of pretest scores for the institution studied. These scores also showed the error was sufficient to cause one to draw incorrect conclusions from the normalized gain. As such, the normalized gain should cease to be used for the purpose of comparing scores for courses with different levels of prior preparation.

RQ2: How can the natural variation of student prior preparation within a course be used to compare classes enrolling students with differing levels of prior preparation? The work above introduced the CGC, multiple methods to calculate a visual representation of the curve, and several linear regression models of the curve. If the CGC is constructed with data drawn from many offerings of a course, then it can be used to evaluate whether a single section of the course performed better or worse than the course as a whole (over the time frame aggregated) correcting for incoming student characteristics. This is done by using either a visual representation of the CGC (Fig. 3) or a mathematical model of the CGC to predict the post-test score from the measured pretest score (and possibly ACTM% scores) of the class section. If this predicted score is below the score actually achieved, the class section outperformed the average course performance corrected for student composition.

RQ3: How can the natural variation of student outcomes within a course be used to evaluate the efficacy of published PER results when transferred to a local context. Using the same method as in the prior research question, the predicted results which would be produced by the local class if it randomly enrolled a class section with the same composition as the used in a published study can be calculated. This could then be compared against the published results to determine if the current course would produce inferior or superior results as those published if it enrolled a student population with a similar composition to those in the published study. We note that this is not quite what would be most useful; ideally we would like to determine what results would be produced if the published course enrolled students of the same composition as the local course. While not optimal, the result that can be calculated would be informative to the decision to elect the published pedagogy.

V. IMPLICATIONS

Following the discussion in RQ 1, normalized gain should not be used to compare conceptual inventory outcomes for institutions with different student populations. It may still be an interesting statistic to characterize educational outcomes; it is simply not reliable for comparing outcomes.

The observation that normalized gain, as well as post-test score, fluctuate with the incoming characteristics of the students implies that it is not appropriate for departments to use uncorrected scores for evaluation purposes or to evaluate the effect of curricular modifications.

Many studies have reported normalized gain and drawn comparisons based on the presumption that they correct for student differences. These studies and conclusions should be revisited. The Hake model in Fig. 3 suggests that the normalized gain may substantially overestimate post-test scores of more weakly prepared student populations while underestimating the score of more prepared students.

The errors produced by using the normalized gain to compare institutions may be having serious negative consequences for the adoption of PER materials at institutions with weaker student populations. Henderson *et al.* showed that adoption of research based methods was very uneven and that methods were often abandoned after they were tried [44]. The course studied in the CL instructional model presents what course personnel believed was a fairly high fidelity implementation of Peer Instruction [39]. The average normalized gain of 36% was viewed as disappointing based on published work and the results of the Hake study [8]. The correction of the 36% to that of the characteristics of the students enrolled at PM in the Salehi *et al.* [15] study produced a normalized gain of 62% which is near the range of high performing programs in Hake's study. The inaccuracy in the normalized gain may cause programs with less well prepared students to underestimate the efficacy of their implementation of PER curricula possibly leading them to discontinue the use of that curricula.

The prediction of the educational efficacy of a PER intervention when implemented for different student population and the comparison of educational outcomes between classes and institutions are crucially important to PER. If the normalized gain does not allow meaningful comparisons between student populations with different characteristics, new methods and statistics must be created for the field to move forward. We propose the reporting and use of CGCs as one such method.

VI. RECOMMENDATIONS

In this work, CGCs were used only for internal evaluation. They allowed two critical questions to be answered: (i) What is the expected conceptual performance of a class section based on the prior preparation of its students? and (ii) What is the expected conceptual performance of a class if it enrolled a student population similar to that in a published PER work? There is a third equally important question which could not be answered: What is the expected conceptual performance of the course in a published work if it enrolled students similar to those in a local course? Answering this question would shed light on whether a published course had particular efficacy for a student population with a particular set of characteristics.

The CGC and the methods used in this work represent one partial solution (the models were not perfect) for comparing conceptual outcomes for students with different characteristics. Until a more efficacious method can be developed, we recommend the following four-step reporting of results. (Later steps are less important than earlier steps).

(1) **Report a summary of the data as fully as possible**: We recommend reporting a table similar to Table I in the Supplemental Material [35] which calculates the post-test average for each possible pretest score. It also includes some additional statistics needed to compute the standard error if a local average is calculated. If the CGC is not reported, it could be calculated from this data.

(2) **Report the conceptual growth curve**: Report the CGC with a representation of the standard error using either binning, a local average, or a more sophisticated smoothing method. All yielded approximately the same results.

(3) **Report a mathematical model of the data**: To allow ease of comparison and to identify the variables most important to the variation in post-test scores, build a mathematical model that accurately (as possible) captures the variation in post-test score with pretest score. If possible, report the additional statistics required to calculate the 95% confidence interval of the curve as explained in the Supplemental Material [35].

(4) **Report a post-test score corrected to a standard value**: None of the above methods allow the simple single-number comparison of outcomes provided by the normalized gain. We propose a post-test score corrected to an incoming ACT/SAT mathematics percentile score of 80% and a pretest score of 35% be reported to allow quick comparisons between studies. The selection of these values is discussed in the next section.

Reporting a post-test score or normalized gain corrected to a standard value is a convenience that allows readers to quickly compare multiple studies. It does not convey the same information about the variation of post-test outcomes with pretest scores as does a CGC; as such the post-test score corrected to a standard value should be reported in addition to the CGC. If a CGC is reported, the actual standard values for pretest score and ACTM% are not particularly important, because the CGC can be used to convert the reported score to any desired pretest score.

VII. SELECTING AND REPORTING A STANDARD VALUE

The standard pretest and ACTM% scores proposed above (35% and 80%, respectively) were selected by examining many studies. Madsen *et al.* [45] reported 11 FCI and 2 FMCE pretest scores disaggregated by gender; FCI pretest scores were somewhat higher than FMCE pretest scores. Examination of the data presented suggests that a representative pretest score of 35% is appropriate for the FCI or FMCE for US universities excluding the most elite institutions. Thornton's *et al.* [33] study also showed FCI scores are somewhat higher than FMCE scores for all but the most prepared students. Many of the institutions reported were very selective and the 35% value was selected weighting more strongly less selective institutions taking into account Kanim and Cid's [46] warning that most PER research has been performed at highly performing institutions and may not be representative of all students. Fewer studies report ACTM% and the 80% value observed in the current study may be reasonable for national comparison.

For the FMCE, the 35% value is within the range of reported scores, but toward the higher side of the range. We note that while many studies use the FMCE, only a relatively small subset of these report descriptive statistics complete enough to infer an overall pretest score. Beyond the pretest scores of the class studied (23%) and of PM in Salehi *et al.* [15] (43.5%), FMCE pretest percentage scores reported include: 24.5% and 42% (2 institutions) [47]; 24% (American students) and 30% (Japanese students) [48]; 30% (one institution) [49], and 16% (one institution) [50].

Using our suggested standard pretest score of 35%, if only pretest score was available, the quadratic model regression equation in Table II predicts a post-test score of $Post = 14.4 + 1.86 \cdot 35 - 0.0104 \cdot 35^2 = 67\%$. The methods in the Supplemental Material [35] allow the calculation of confidence intervals; these methods produce a 95% post-test confidence interval of $Pre = 35\%$ is [65.6%, 67.8%]. Using the quadratic model with ACTM % in Table II at $Pre = 35\%$ and ACTM% = 80% yields a post-test score of 65% and a 95% confidence interval for the post-test of [64.0%, 66.0%].

VIII. LIMITATIONS

This study was performed at a single institution; we expect similar studies at different institutions to produce different results. As such, the results should not be considered as general. In fact, as the variation between institutions is reported, we expect important new insights about physics instruction to emerge.

The primary variable accounting for the variance in post-test score in this study was pretest score. The study also reported models using standardized test scores to improve predictions. Standardized test scores have often been used as control variables in PER studies which resulted in their inclusion in the present study. As more institutions move away from requiring the reporting standardized test scores to support holistic admission policies, these measures may become less commonly available. It is possible in the future that PER studies move to another control variable such as high school grade point average (HSGPA). In anticipation of this eventuality, the Supplemental Material [35] present the CGC disaggregated by HSGPA, the by-semester plot of HSGPA, and the quadratic regression model predicting post-test from pretest and HSGPA. HSGPA does not have the same explanatory power of ACTM% explaining only 1% additional variance over pretest alone.

Although the conceptual understanding of mechanics measured by FMCE scores is an important outcome of introductory physics classes, it is far from the only goal of such courses. The FMCE fails to capture students' growth in areas such as mathematical sophistication and independent problem solving, and, as such, should not be used as the only metric for evaluating instructional success.

This study did not use demographic variables such as gender or first-generation college student status. Gender has long been identified as an important variable in predicting pretest and post-test scores [45]. Gender predicts about 3% of the variance in post-test scores when added to a model containing pretest and ACTM%, much less than pretest score and about half the variance of ACTM%. We chose not to include it in the analysis or the recommendations for comparison between institutions. Until the source of these gender differences is identified, it seemed irresponsible to suggest post-test scores be corrected for gender composition.

This study investigated the results of the FMCE and discussed the FCI. Recently, quantitative analysis has shown these now venerable instruments have significant flaws. Factor analysis has shown neither instrument has a useful factor structure [29,50–56]; the factor structure extracted for the FCI differs from that published with the instrument. Classical test theory has been used to identify some items in each instrument which are problematic [57,58]. Differential item functioning theory identified many items in the FCI which were unfair to either men or women as well as a few unfair items in the FMCE [57,58]. As such, while the present study focuses on the FCI and FMCE as examples, it is our hope these instruments will soon be revised and the CGC will be used to report the results of applying these new instruments.

IX. FUTURE WORK

The wealth of FCI and FMCE data collected since the introduction of the instruments opens the possibility of computing the CGCs for many institutions implementing a variety of pedagogies. This would allow the determination of instructional techniques which best support students with different levels of academic preparation allowing the targeting of the most effective instructional methods for all students. This work will also be extended to demographic groups underrepresented in physics classes to determine if physics classes are reaching all students and to identify pedagogies that equitably serve all students.

X. CONCLUSIONS

FMCE pretest, post-test, and ACT/SAT percentile scores varied by semester at the institution studied. Neither the normalized gain, nor Cohen's *d* removed this variation. The local average of a graph of post-test scores plotted against pretest scores provided a more detailed characterization of the conceptual outcomes of the class for students with differing incoming preparation in physics. Such a plot could be included in educational studies to facilitate comparison of educational innovations across institutions and instructional models; one could also report an approximate mathematical model of the curve. If a single metric for comparison is desired, correcting results to a standardized pretest score would allow accurate comparison between institutions. Predicting the results of the class studied using the incoming characteristics of students in a published PER work dramatically changed the interpretation of the relative effectiveness of the class studied, showing such correction is crucial to understanding the general impact of pedagogical methods.

The variation observed in student preparation in the class studied implies that the conceptual inventory outcomes of the class cannot directly be used to compare instructors or educational reforms. Graphical or mathematical models which incorporate the student variation should be constructed before using conceptual inventory scores (or other outcome metrics) to evaluate differences in instruction.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grants No. ECR-1561517, No. HRD-1834569, and No. DUE-1833694. We would like to thank the referees for their insights; the work evolved substantially for the better under review.

[1] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[2] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[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] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).

[5] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).

[6] D. E. Meltzer and R. K. Thornton, Resource letter ALIP-1: Active-learning instruction in physics, Am. J. Phys. **80**, 478 (2012).

[7] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).

[8] 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).

[9] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 K-student study, Am. J. Phys. **84**, 969 (2016).

[10] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

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

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

[13] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).

[14] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting Force Concept Inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).

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

[16] J. Stewart, G. L. Cochran, R. Henderson, C. Zabriskie, S. DeVore, P. Miller, G. Stewart, and L. Michaluk, Mediational effect of prior preparation on performance differences of students underrepresented in physics, Phys. Rev. Phys. Educ. Res. **17**, 010107 (2021).

[17] D. Hewagallage, E. Christman, and J. Stewart, Examining the relation of high school preparation and college achievement to conceptual understanding, Phys. Rev. Phys. Educ. Res. **18**, 010149 (2022).

[18] Lei Bao, Theoretical comparisons of average normalized gain calculations, Am. J. Phys. **74**, 917 (2006).

[19] V. P. Coletta and J. J. Steinert, Why normalized gain should continue to be used in analyzing preinstruction and post-instruction scores on concept inventories, Phys. Rev. Phys. Educ. Res. **16**, 010108 (2020).

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

[21] A. Madsen, S. B. McKagan, and E. C. Sayre, Best practices for administering concept inventories, Phys. Teach. **55**, 530 (2017).

[22] C. Vieira, P. Parsons, and V. Byrd, Visual learning analytics of educational data: A systematic literature review and research agenda, Comp. Educ. **122**, 119 (2018).

[23] T. Martin-Blas, L. Seidel, and A. Serrano-Fernández, Enhancing Force Concept Inventory diagnostics to identify dominant misconceptions in first-year engineering physics, Eur. J. Eng. Educ. **35**, 597 (2010).

[24] M. D. Caballero, E. F. Greco, E. R. Murray, K. R. Bujak, M. Jackson Marr, R. Catrambone, M. A. Kohlmyer, and M. F. Schatz, Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study, Am. J. Phys. **80**, 638 (2012).

[25] G. Poutot and B. Blandin, Exploration of students' misconceptions in mechanics using the FCI, Am. J. Educ. Res. **3**, 116 (2015).

[26] Y. Shoji, S. Munejiri, and E. Kaga, Validity of Force Concept Inventory evaluated by students' explanations and confirmation using modified item response curve, Phys. Rev. Phys. Educ. Res. **17**, 020120 (2021).

[27] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

[28] S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch, and N. R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison, Eur. J. Phys. **34**, 421 (2013).

[29] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory dataset, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).

[30] J. Stewart, B. Drury, J. Wells, A. Adair, R. Henderson, Y. Ma, A. Pérez-Lemonche, and D. Pritchard, Examining the relation of correct knowledge and misconceptions using the nominal response model, Phys. Rev. Phys. Educ. Res. **17**, 010122 (2021).

[31] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, Am. J. Phys. **74**, 449 (2006).

[32] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80**, 825 (2012).

[33] 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).

[34] L. E. Kost-Smith, S. J. Pollock, and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a "smog of bias", Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).

[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.20.010123 for conceptual growth curves with error bars, the data needed to calculate 95% confidence intervals for the regression models, and a discussion of the statistics of computing these intervals.

[36] National Center for Education Statistics, https://nces.ed.gov/collegenavigator. Accessed 1/30/2022.

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

[38] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Prentice Hall, Upper Saddle River, NJ, 1998).

[39] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).

[40] E. Elby, R. E. Scherr, T. McCaskey, R. Hodges, T. Bing, D. Hammer, and E. F. Redish, Open Source Tutorials in Physics Sensemaking, http://umdperg.pbworks.com/w/page/10511218/Open Source Tutorials. Accessed 9/17/2018.

[41] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, NY, 1977).

[42] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer-Verlag, New York, NY, 2003).

[43] R. McElreath, *Statistical Rethinking: A Baysian Course with Examples in R and Stan* (CRC Press, Taylor & Francis Group, Boca Raton, FL, 2016).

[44] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, Phys. Rev. ST Phys. Educ. Res. **8**, 020104 (2012).

[45] A. Madsen, S. B. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

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

[47] R. Henderson, J. Stewart, and A. Traxler, Partitioning the gender gap in physics conceptual inventories: Force Concept Inventory, Force and Motion Conceptual Evaluation, and Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **15**, 010131 (2019).

[48] M. Ishimoto, G. Davenport, and M. C. Wittmann, Use of item response curves of the Force and Motion Conceptual Evaluation to compare Japanese and American students' views on force and motion, Phys. Rev. Phys. Educ. Res. **13**, 020135 (2017).

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

[50] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, Am. J. Phys. **76**, 882 (2008).

[51] J. Yang, C. Zabriskie, and J. Stewart, Multidimensional item response theory and the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **15**, 020141 (2019).

[52] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, Phys. Teach. **33**, 138 (1995).

[53] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[54] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).

[55] M. R. Semak, R. D. Dietz, R. H. Pearson, and C. W. Willis, Examining evolving performance on the Force Concept Inventory using factor analysis, Phys. Rev. Phys. Educ. Res. **13**, 010103 (2017).

[56] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force

Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010137 (2018).

[57] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

[58] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **14**, 020103 (2018).