AAPT PHYSICS EDUCATION | AMERICAN JOURNAL of PHYSICS

Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study

Marcos D. Caballero, Edwin F. Greco, Eric R. Murray, Keith R. Bujak, M. Jackson Marr, Richard Catrambone, Matthew A. Kohlmyer, and Michael F. Schatz

Citation: American Journal of Physics **80**, 638 (2012); doi: 10.1119/1.3703517
View online: http://dx.doi.org/10.1119/1.3703517
View Table of Contents: http://scitation.aip.org/content/aapt/journal/ajp/80/7?ver=pdfcov
Published by the American Association of Physics Teachers

PHYSICS EDUCATION RESEARCH SECTION

The Physics Education Research Section (PERS) publishes articles describing important results from the field of physics education research. Manuscripts should be submitted using the web-based system that can be accessed via the American Journal of Physics home page, http://ajp.dickinson.edu, and will be forwarded to the PERS editor for consideration.

Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study

Marcos D. Caballero,[a] Edwin F. Greco, and Eric R. Murray
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332

Keith R. Bujak, M. Jackson Marr, and Richard Catrambone
School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332

Matthew A. Kohlmyer
Advanced Instructional Systems, Inc., Raleigh, North Carolina 27696

Michael F. Schatz[b]
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332

(Received 6 July 2011; accepted 29 March 2012)

The performance of over 5000 students in introductory calculus-based mechanics courses at the Georgia Institute of Technology was assessed using the Force Concept Inventory (FCI). Results from two different curricula were compared: a traditional mechanics curriculum and the Matter & Interactions (M&I) curriculum. Both were taught with similar interactive pedagogy. Post-instruction FCI averages were significantly higher for the traditional curriculum than for the M&I curriculum; the differences between curricula persist after accounting for factors such as pre-instruction FCI scores, grade point averages, and SAT scores. FCI performance on categories of items organized by concepts was also compared; traditional averages were significantly higher in each concept. We examined differences in student preparation between the curricula and found that the relative fraction of homework and lecture topics devoted to FCI force and motion concepts correlated with the observed performance differences. Concept inventories, as instruments for evaluating curricular reforms, are generally limited to the particular choice of content and goals of the instrument. Moreover, concept inventories fail to measure what are perhaps the most interesting aspects of reform: the non-overlapping content and goals that are not present in courses without reform.
[http://dx.doi.org/10.1119/1.3703517]

I. INTRODUCTION

Each year more than 35% of American college and university students enroll in a physics course.[1] Only a small fraction of these students ultimately complete a degree in physics; the vast majority pursue a degree in engineering or another science.[2] Many are students in an introductory physics course; approximately 175,000 students each year enroll in introductory calculus-based physics.[3] However, many of these students fail to acquire an effective understanding of concepts, principles, and methods from these introductory courses. Rates of failure and withdrawal from these courses are often high and substantial research into this subject has shown that students' misconceptions in physics persist after instruction.[4,5] This paper describes an attempt to evaluate, using a multiple-choice concept inventory,[6] a reformed introductory mechanics curriculum[7] which aims to mitigate these issues by altering the goals and content (i.e., the **curriculum**) of the typical mechanics course.

To help improve student learning in physics, many new methods of content delivery (**pedagogy**) have been developed in recent years. Typically, these methods have been implemented with a little change to course curricula. Well established pedagogical modifications now used widely include tutorials,[8] clicker questions,[9] peer instruction,[10] Socratic tutorial homework systems,[11] multiple representations of concepts and principles,[12] and reconfigurations of the instructional environment.[13] There is ample evidence that students who experience these pedagogical reforms perform better on end-of-course concept inventories than students in passive lecture courses. Concept inventories are useful tools to make such comparisons in these cases where all courses (with and without pedagogical reform) share, for the most part, the same core curriculum.

By contrast, there is sparse research on how student learning is affected by substantial alterations to the curriculum of introductory physics courses. One reason for the lack of such work is the relative absence of alternative introductory physics curricula; improvements to the introductory physics curriculum have not progressed as rapidly as improvements in pedagogy. Most students learn introductory physics following a canon of topics that has remained largely unchanged for

decades regardless of the textbook edition or authors. Moreover, choosing how to compare a novel introductory physics curriculum to a traditional curriculum presents a challenge. Concept inventories can be used for such a comparison.[14] However, there are a number of issues that are peculiar to curricular comparison including which topics to select for comparison and the alignment of the inventory with the goals and content of the curricularly reformed course.

At the Georgia Institute of Technology (Georgia Tech, GT), we have used a concept inventory to evaluate student understanding of force and motion in both a traditionally sequenced introductory calculus-based mechanics course[15] and an introductory calculus-based mechanics course using the reform curriculum, Matter & Interactions (M&I).[7] While both courses employ similar pedagogical best practices, M&I differs from the traditional curriculum in its focus on the generality of fundamental physical principles, the introduction of microscopic models of matter, and its coherence in linking different domains of physics.[16,17] In particular, M&I revises the learning progression of the first semester introductory mechanics course by reorganizing and augmenting the traditional sequence of topics. For example, early emphasis is placed on the impulse-momentum theorem (referred to as the "momentum principle" in the M&I curriculum), $\Delta\vec{p} = \vec{F}\Delta t$, with iterative application of the momentum principle over short time steps to predict motion by means of both analytic calculation and numerical computation.[18,19] Furthermore, M&I introduces non-constant forces early on to demonstrate the predictive power of this principle. By contrast, in a traditional curriculum, early emphasis is placed on study of the kinematics of special case situations (e.g., motion under constant acceleration) without explicit discussion of dynamics. Further discussion of differences between the M&I curriculum and a traditional curriculum can be found elsewhere.[7,16–18]

At present, there is no mechanics concept inventory, whose force and motion content has been explicitly aligned with goals and content in both courses both traditional and M&I mechanics reform. (By contrast, there is at least one concept inventory that is aligned with both traditional and M&I electromagnetism).[14,20] Under these circumstances, we chose the Force Concept Inventory (FCI) to make a comparative evaluation both because it is widely used and because of anecdotal evidence of underperformance on the FCI in courses using M&I mechanics at other institutions. The FCI was designed to probe performance on force and motion in a particular way; within the context of specific situations, FCI questions were designed to draw out common misconceptions and naive notions about force and motion.[6] As a result, a measurement of performance using the FCI does not provide a comprehensive picture of student understanding of force and motion; the nuances of interpreting student performance on the FCI have been well-documented.[21–25] Furthermore, the FCI was not specifically developed to compare student performance between courses but has been used for this purpose.[26] Thus, to emphasize the idea that the FCI probes force and motion in a restricted way, we indicate, in this paper, the content of and concepts covered by the FCI as *FCI force and motion concepts* (To obtain a copy of the FCI, contact David Koch (ASU) by email: FCIMBT@verizon.net). Moreover, we qualify all of our comparative measures with the understanding that the FCI was designed in the context of a traditional sequenced curriculum before the M&I curriculum came into existence.

The description of our study is presented below as follows: In Sec. II, we describe the organizational structure of the Georgia Tech mechanics courses. Section III summarizes the results of the in-class testing. In Sec. IV, we present an analysis of FCI performance by individual item and concept. Section V examines possible reasons for performance differences observed in Secs. III and IV. In Sec. VI, we provide more insight into the performance differences, make concluding remarks, and outline possible future research directions.

II. INTRODUCTORY MECHANICS AT GEORGIA TECH

The typical introductory mechanics course at Georgia Tech is taught with three one-hour lectures per week in large lecture sections (150–250 students per section) and three hours per week in small group (20 student) laboratories and/or recitations. Attendance of lecture sections is optional but encouraged through a small incentive (2–5% of course grade). Attendance of laboratory and recitation sections is mandatory. In the traditional (TRAD) curriculum, each student attends a 2 h laboratory and, in a separate room, a 1 h recitation each week (the use of the label "traditional" to describe the non-M&I course is a matter of convenience. Georgia Tech's traditional course is pedagogically reformed, however, the traditional course explores the typical content and examples presented in most introductory physics courses). In the M&I curriculum, students meet once per week with teaching assistants for a single 3 h laboratory/recitation session involving both lab activities (for approximately 2 h on average) and separate recitation activities (for approximately 1 h on average). Room/TA scheduling is responsible for the differences in instructional locations between the two courses. The student population of the mechanics course (both traditional and M&I) consists of approximately 85% engineering majors and 15% science (including computer science) majors.

Table I summarizes the FCI test results for individual sections. In most traditional (T6–T22) and all M&I sections, N_O students in each section took the FCI during the last week of class at the completion of the course. In all of the traditional sections and in the majority of M&I sections (M2–M6), N_I students in each section took the FCI at the beginning of the course during the first week of class. For a given section, N_I is approximately equal to the number of students enrolled in that section. N_O is usually smaller than N_I, sometimes substantially so (e.g., T12, T13, and T20). M&I students took both the pre- and post-test during their required laboratory section. Students of the traditional curriculum typically took the pre-test during the first lecture or lab section. Traditional students were asked to attend an optional section during their evening testing period to take the post-test. Students become busy with other coursework near the end of the semester, hence fewer traditional students attended this optional evening section. In each section, only those N_m students who took the FCI both on entering and on completion of the course are considered for the purposes of computing any type of gain (Sec. III). The FCI was administered using the same time limit (30 min) for both traditional and M&I students. M&I students were given no incentives for taking the FCI; they were asked to take the exam seriously and told that the score on the FCI would not affect their grade in the course. Traditional students taking the FCI were given bonus credit worth up to a maximum of 0.5% of their final course

Table I. Georgia Tech FCI test results are shown for 22 traditional sections (T1–T22) and 6 Matter & Interactions sections (M1–M6). Different lecturers are distinguished by a unique letter in column L. The average incoming FCI score I for N_I students entering the course is shown for sections in which the FCI was given prior to instruction. In those sections where data are available, the average outgoing FCI score O for N_O students completing this course are indicated. N_m is the number of students in a given section who took the FCI both at the beginning and at the end (i.e., matched data) of their mechanics course.

ID	L	I%	N_I	O%	N_O	N_m
T1	A	49.95 ± 3.05	194	N/A	N/A	N/A
T2	A	52.13 ± 2.80	208	N/A	N/A	N/A
T3	B	51.76 ± 2.88	207	N/A	N/A	N/A
T4	B	51.39 ± 2.91	196	N/A	N/A	N/A
T5	C	46.39 ± 2.69	205	N/A	N/A	N/A
T6	D	45.83 ± 3.53	139	70.13 ± 3.60	103	97
T7	C	47.27 ± 2.86	182	64.01 ± 3.05	158	139
T8	C	42.03 ± 2.55	194	61.26 ± 3.14	140	133
T9	A	52.16 ± 2.99	182	73.44 ± 2.97	127	122
T10	A	48.12 ± 2.72	188	73.97 ± 2.92	116	113
T11	B	49.82 ± 2.88	182	75.35 ± 3.48	104	98
T12	B	49.58 ± 3.43	168	72.04 ± 4.06	93	88
T13	E	52.81 ± 3.25	141	77.20 ± 3.38	88	84
T14	E	40.36 ± 2.65	183	67.33 ± 3.53	140	132
T15	F	46.39 ± 3.05	180	69.59 ± 3.36	131	120
T16	F	40.74 ± 2.84	194	65.22 ± 3.60	115	108
T17	E	48.02 ± 3.17	160	71.82 ± 3.57	121	109
T18	A	50.19 ± 3.05	175	74.05 ± 3.44	107	105
T19	A	53.49 ± 3.37	174	72.10 ± 3.52	103	94
T20	E	53.36 ± 3.27	143	78.52 ± 3.68	97	89
T21	B	49.43 ± 3.00	180	75.79 ± 3.12	121	115
T22	B	51.48 ± 3.09	182	79.92 ± 2.81	119	116
M1	G	N/A	N/A	35.71 ± 5.62	28	N/A
M2	H	54.12 ± 3.86	127	64.68 ± 4.16	116	111
M3	G	45.01 ± 3.11	145	56.49 ± 3.38	148	133
M4	H	45.57 ± 3.51	143	62.27 ± 3.37	141	128
M5	I	45.35 ± 3.61	134	62.70 ± 3.44	132	110
M6	J	44.83 ± 2.50	214	54.15 ± 3.06	196	180

score, depending in part on their performance on the FCI. This incentive difference between the two curricula has no bearing on the performance differences we observe in our data (Sec. V).

III. SUMMARY OF MEASUREMENTS FROM IN-CLASS TESTING

The FCI pre-test scores for Matter & Interactions (M&I) and traditional students did not differ significantly (mean FCI score of 48.9% for TRAD vs 47.4% for M&I). By contrast, on the FCI post-test, traditional students significantly outperformed M&I students (mean FCI score of 71.3% for TRAD vs 59.3% for M&I). In Fig. 1, these mean scores have been reported with 95% confidence intervals estimated from the t-statistic for each distribution.[27] A common measure of the change in performance from pre-test to post-test[26] is the average percentage gain, $G = (O - I) * 100\%$, where I is the average fractional FCI score for students entering a mechanics course, and O is the average end-of-course fractional FCI score. We also report an average normalized gain g, where $g = (O - I)/(1 - I)$, and where $(1 - I)$ represents the maximum possible fractional gain that could be obtained by a

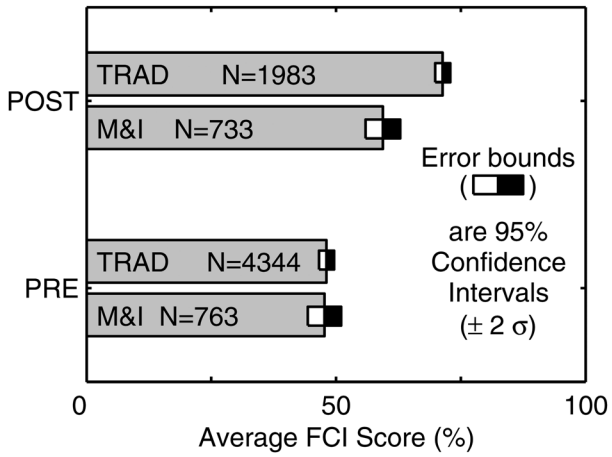

Fig. 1. Average pre- and post-instruction FCI scores at Georgia Tech. The average FCI pre- and post-test scores are shown for students who took a one-semester mechanics course with either the traditional (TRAD) or Matter & Interactions (M&I) curriculum. The number of students (N) tested for each curriculum is indicated in the figure. The error bounds represent the 95% confidence intervals (estimated from the t-statistic) on the estimate of the average score.

class of students with an average incoming fractional FCI score of I. For the gains reported in Fig. 2, 95% confidence intervals have been estimated from the t-statistic for the distributions of G and g. The data are shown for N_m students (Table I).

FCI pre-test score distributions were found to be statistically indistinguishable between the two curricula, which is evident from Fig. 3(a). By contrast, distributions of post-test FCI scores were dissimilar; the traditional distribution was shifted towards higher scores [Fig. 3(b)]. This is consistent with the finding that the mean score achieved by traditional students were higher than their M&I peers on the post-test (Fig. 1). Because the distributions of FCI pre- and post-test scores were non-normal, the similarity of the distributions was compared using a rank-sum test.[28,29]

An examination of measures of student performance entering each course suggests that the incoming and

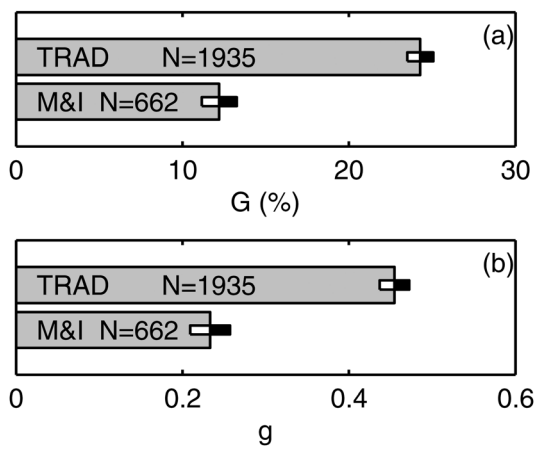

Fig. 2. Gain in understanding of mechanics as measured by the FCI. The increase in student understanding resulting from a one-semester traditional (TRAD) or Matter & Interactions (M&I) course is measured using (a) the average raw gain G and (b) the average normalized gain g. Only students with matched scores were used for this figure (see Table I). The error bounds represent the 95% confidence intervals (estimated from the t-statistic) on the estimate of (a) the raw gain and (b) the normalized gain.

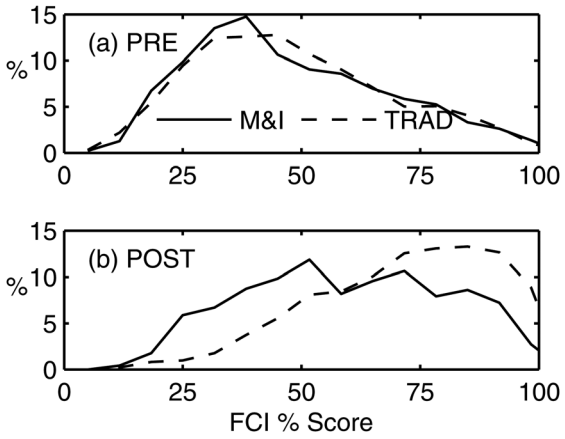

Fig. 3. FCI score distributions by curriculum. The distributions of FCI test scores for students before (a) and after (b) completing a mechanics course with either a traditional (dashed line) or M&I curriculum (solid line) are shown. The total number of students tested in each curriculum is the same as in Fig. 1. The plots are constructed from binned data with bin widths equal to approximately 6.7% of the maximum possible FCI score (100%).

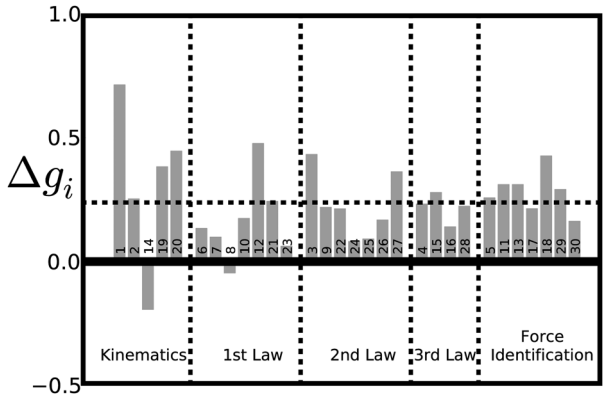

Fig. 4. Difference in performance for individual FCI items and mechanics concepts. The difference in performance Δg_i between traditional and M&I students is shown for each question on the FCI. Positive (negative) Δg_i indicates superior performance by traditional (M&I) students on individual questions. The numerical labels indicate the corresponding question number in order of appearance on the FCI. The items are grouped together into one of five concepts: Kinematics, Newton's first law, Newton's second law, Newton's third law, and Force Identification. The horizontal line (dash) illustrates the value of $\overline{\Delta g}$, the mean difference in the item gains between curricula.

outgoing student populations of both curricula were identical. We obtained and examined students' grade point averages (GPA) upon entering the mechanics course, SAT Reasoning Test (SAT) scores, and the grades earned in the mechanics course; we found no significant difference in the distributions of any of these metrics using a rank-sum test.

Mean scores differed between one or more sections within a given curriculum as measured by a Kruskal-Wallis test.[31] Given this section effect, we compared the three lowest performing traditional sections (T7, T8, and T16) to the three highest performing M&I sections (M2, M4, and M5) to determine if this section effect enhanced the overall observed differences in the normalized gains. Post-test FCI scores were statistically indistinguishable between these subsets. However, traditional students in these sections had significantly lower pre-test FCI scores. Hence, students in these lower performing traditional sections achieved significantly higher normalized gains. We also compared the FCI post-test scores achieved by the three traditional sections with lowest normalized gains (T14, T18, and T22) to the M&I sections with the highest normalized gains (M3, M4, and M5). Pre-test FCI scores were significantly higher for the M&I subset, while post-test scores were higher for the traditional subset. Thus normalized gains achieved by traditional students in this subset were higher.

IV. ITEM ANALYSIS OF FCI MEASUREMENTS

Student performance on individual questions or groups of questions was used to determine on which FCI force and motion concepts students in the traditional curriculum outperformed M&I students. Questions on the FCI were sorted into concept categories using Hestenes' original conceptual dimensions,[6] but we required that each question be placed in only one category. In our work, only five concept categories were used: Kinematics, Newton's first Law, Newton's second Law, Newton's third Law, and Force Identification. The first four of these categories were identical to Hestenes' dimensions and Force Identification was a renamed category which contained questions from Hestenes' Kinds of Forces dimension. In Fig. 4, the items that comprise each category are listed. Note that this was an *a priori* categorization based on our judgment of the concepts covered by the items; it is not the result of internal correlations or factor analysis based on student data.

We used the normalized gain in performance on a per question basis to quantify item performance. We define an item gain,

$$g_i = \frac{f_{\text{post},i} - f_{\text{pre},i}}{1 - f_{\text{pre},i}}, \tag{1}$$

where $f_{\text{pre},i}$ and $f_{\text{post},i}$ are the fraction of students responding correctly to the ith item on the pre- and post-test, respectively. This measure normalizes the gain in performance on a single item by the largest possible gain given the students' pre-test performance on that item; g_i is essentially the Hake gain for a single item. To discern which questions have large item gains, we can compare g_i for each question to the mean item gain,

$$\bar{g} = \frac{1}{N}\sum_{i}^{N} g_i, \tag{2}$$

where N is the number of items on the FCI.

To illustrate the differences between curricula succinctly, we computed the difference in normalized item gains between the two curricula. We define the difference in normalized item gain,

$$\Delta g_i = g_i^T - g_i^M, \tag{3}$$

where g_i^T and g_i^M are the normalized gain for the ith item achieved by traditional and M&I students, respectively. We discovered on which questions students' item gains in each curriculum differed the most by comparing Δg_i for each item to the mean difference in the item gains between curricula,

$$\overline{\Delta g} = \frac{1}{N}\sum_{i}^{N} \Delta g_i. \tag{4}$$

The plot of Δg_i illustrates better performance by traditional students across all concepts on the FCI (Fig. 4). We observed

that Δg_i is positive for almost all questions, and 45% of the questions had values of Δg_i greater than $\overline{\Delta g} = 0.238$. The grouping of the FCI questions by category permits one to visualize which concepts contributed most strongly to the difference in performance. For example, the difference in performance on the Force Identification concept was striking, where five of the seven questions in this category had $\Delta g_i > \overline{\Delta g}$.

Moreover, this grouping helps one to determine on which concepts differences in item gains were greatest. We computed the difference in the average concept gain,

$$\overline{\Delta g_c} = \frac{1}{N_c} \sum_{i \epsilon c} \Delta g_i, \tag{5}$$

where N_c is the number of items covering concept c. Concepts with higher $\overline{\Delta g_c}$ were those on which traditional students achieved higher normalized gains than M&I students. The Kinematics and Force Identification concepts had the highest values of $\overline{\Delta g_c}$ (shown in Table II). By contrast, we found $\overline{\Delta g_c}$ for Newton's first law which was well below $\overline{\Delta g}$. The remaining two concepts had values of $\overline{\Delta g_c}$ slightly below $\overline{\Delta g}$.

V. CONTRIBUTIONS TO THE PERFORMANCE DIFFERENCES

We turn now to the examination of factors that might contribute to higher FCI post-test scores by traditional students, including grade incentives, differences in pedagogy, and differences in instruction (e.g., homework and lecture topics).

The incentive given to traditional students to take the FCI was too small to account for the marked differences in performance indicated in Figs. 1, 2, and 3(b). As mentioned earlier (Sec. II), traditional students were provided with an incentive to take the FCI, while M&I students received no incentive. In principle, sufficiently large incentives can impact FCI outcomes. For example, Ding *et al.*, found a 10–15% increase in FCI post-test scores if scores on the FCI were valued as highly as course exams.[30] To check for this incentive effect, we offered similar incentives (i.e., a maximum of 0.5% bonus to overall course grade) to both traditional and M&I students who took the FCI post-test at Georgia Tech in the fall of 2009. During this term, we found the performance differences for M&I and traditional students were similar to those reported in this paper. FCI data from fall 2009 were not included in this paper because instructional changes had been made to the M&I course; M&I sections M1–M5 had similar homework exercises, lectures, and laboratories.

The performance differences cannot be attributed to differences in pedagogy. It is well known that using interactive engagement (i.e., "clicker" questions, ConceptTests, Peer Instruction, etc.) can improve students' conceptual understanding in introductory and advanced courses.[26,31,32] However, all sections (both traditional and M&I) were largely indistinguishable with respect to interactive engagement: all sections used similar methods ("clicker" questions) with similar intensity (3–6 "clicker" questions per lecture period).

We examined whether differences in coursework (homework) could be connected to performance differences on the FCI. We categorized the 575 traditional homework questions and the 756 M&I homework questions. Questions were placed into one or more categories depending on the topical nature of the problem and the principles needed to answer the question. Categories included the five FCI force and motion concepts discussed in Sec. IV as well as several other concepts which do not appear on the FCI (e.g., angular momentum). The *Kinematics category* included questions about the relationships between position, velocity, and acceleration that did not refer to the underlying dynamical interactions that cause changes in these quantities. Questions in the *Newton's first law category* included qualitative questions which discussed the direction of motion and its relationship to applied forces. The *Newton's second law category* included questions with a heavy emphasis on contact forces and resolving unknown forces, but excluded open-ended questions in which the prediction of future motion is the goal (e.g., using iterative methods to predict the motion of an object). Questions in the *Newton's third law category* included those in which Newton's third law was treated as an isolated law, that is, where there was no reference to the underlying reciprocity of long-range electric interactions which causes it. Generally, it was applied to contact forces and gravitational interactions. The *Force Identification category* included questions in which the direction and relative strength of forces acting on a body or set of bodies were represented by diagrams (i.e., force-body diagrams). The aforementioned categories represent those concepts that are covered extensively in the first half of a traditional physics course and were heavily represented on the FCI.

The difference in the relative fraction of homework questions covering FCI force and motion concepts between the curricula (Table III) reflect the overall performance

Table II. The average differences in item gains between curricula are computed for the items in each FCI force and motion concept, $\overline{\Delta g_c}$. Each $\overline{\Delta g_c}$ is positive, indicating better average item gains for traditional students across all FCI force and motion concepts. Concepts with higher $\overline{\Delta g_c}$ are those for which traditional students achieve higher normalized gains than M&I students. Traditional students achieve the highest values of $\overline{\Delta g_c}$ on the Kinematics and Force Identification concepts and lowest on Newton's first law concept. The measures are presented along with their variance.

FCI force and motion concepts	$\overline{\Delta g_c}$	σ^2
Kinematics	0.32	<0.01
Newton's first law	0.16	<0.01
Newton's second law	0.22	<0.01
Newton's third law	0.22	0.01
Force identification	0.28	0.05

Table III. An estimate of the fraction of homework questions covering a particular FCI concept in the two mechanics curricula is compared. Subtopics for these homework questions were not mutually exclusive. The relative fraction of homework questions covering FCI force and motion concepts and some individual FCI concepts (i.e., Kinematics, Newton's second law, Newton's third law, and Force Identification) is greater in the traditional curriculum.

Est. Fraction of HW Questions	M&I	TRAD
FCI force and motion concepts	0.26	0.57
HW Subtopics (not exclusive)		
Kinematics	0.10	0.26
Newton's first law	<0.01	<0.01
Newton's second law	0.15	0.25
Newton's third law	<0.01	0.04
Force identification	0.01	0.11

Table IV. Comparison of the estimated fractions of lecture/reading topics in the two mechanics curricula. Subtopics for these lectures/readings were not mutually exclusive. The relative fraction of lectures/readings in the traditional course is greater for the Kinematics, Newton's third law, and Force Identification topics, which is consistent with their superior performance in those concepts on the FCI. However, on Newton's first and second laws, the relative fractions of lectures/readings are roughly similar.

Estimated fraction of lecture topics	M&I	TRAD
FCI force and motion concepts	0.26	0.44
Lecture subtopics (not exclusive)		
Kinematics	0.07	0.21
Newton's first law	0.02	0.01
Newton's second law	0.09	0.08
Newton's third law	0.01	0.03
Force identification	0.06	0.11

differences observed in Figs. 1–3. Furthermore, the differences in the relative fractions of homework questions corresponding to individual FCI concepts were consistent with the results from our item analysis (Fig. 4). The relative fraction of homework questions was computed by first categorizing questions, then counting the number of questions covering the concepts of interest and dividing by the total number of homework questions given in a curriculum. The relative fraction of homework questions covering FCI force and motion concepts differed by more than a factor of 2 in favor of the traditional curriculum. On individual FCI concepts, we found a lower relative fraction of homework questions in the M&I curriculum compared with the traditional curriculum on four of the five concepts: Kinematics, Newton's second law, Newton's third Law, and Force Identification. On most FCI questions about these concepts traditional students outperformed M&I students (Sec. IV and Fig. 4). We found that the relative fraction of Newton's first law questions were similar. This signature was also observed in our item analysis (Fig. 4); the Newton's first law FCI concept had the smallest $\overline{\Delta g_c}$ (Sec. IV).

The difference in the relative fraction of force and motion lectures/readings between the curricula (Table IV) was consistent with the overall performance differences observed in Figs. 1, 2, and 3(b). The relative fraction of lectures/readings which cover FCI force and motion concepts was greater by nearly a factor of 2 for the traditional curriculum. This result is consistent with the difference in the relative fraction of homework questions (Table III). However, the differences in the relative fractions of lectures/readings corresponding to individual FCI concepts showed mixed results when compared to our item analysis (Fig. 4). The relative fractions for three of five concepts were greater for the traditional curriculum: Kinematics, Newton's third law, and Force Identification. But on two concepts, the relative fractions of lectures/readings were roughly similar: Newton's first law and Newton's second law. Lecture and reading topics were examined and categorized for each curriculum using the same categories as our homework question analysis.

VI. CLOSING REMARKS AND LESSONS LEARNED

We have found that students who completed an introductory mechanics course which employs the Matter & Interactions curriculum earned lower post-test FCI scores than students who took a traditional curriculum. The differences in performance were significant, given the large number of students involved in the measurement. We demonstrated that these differences cannot be explained by differences in the incoming or outgoing population of students between the courses (i.e., SAT scores, GPA, etc.). The overall performance difference between the curricula on the post-test was consistent with the substantial difference in the amount of directly applicable instruction within each curriculum. The relative fraction of FCI force and motion concepts that appeared on students' homework and in their lectures was roughly twice as large for the traditional curriculum (Tables III and IV). We observed this signature in the differences of the means and distributions of FCI scores [Figs. 1, 2, and 3(b)] as well as the average item gain [Eq. (2)]. The average item gain for traditional students was roughly twice as large as that of M&I students (Sec. IV). Furthermore, we found that traditional students outperformed M&I students across all subtopics on the FCI (Fig. 4) and that these differences were consistent with the amount of instruction on individual FCI force and motion concepts that appeared on students' homework (Table III).

These results indicate the challenges that arise when concept inventories are used to make comparative evaluations of curricular course reforms. Such challenges, it should be emphasized, do not typically arise when concept inventories are used to evaluate pedagogical reforms, which often do not affect core course content. There are at least two considerations that must be kept explicitly in mind for the case of curricular reform. First, sensible comparison between courses with and without reform can be made only on content that is present both in courses. Comparing student performance on curricular materials exclusive to one or the other course (e.g., computation in the case of M&I mechanics) makes little sense. Substantial content on force and motion is found in both traditional and M&I curricula; however, as was mentioned in Sec. I, the specifics of force and motion content differ substantially between the two curricula. Second, the composition of the evaluation instrument itself represents a particular selection of content and goals. Ideally, for a comparative evaluation, the content in the instrument should be aligned with content present in both courses (with and without reform) under study; moreover, the goals evaluated by the instrument should be clearly connected to the learning goals of both courses. Our results support the idea that the content of the FCI is more closely aligned with the content of traditional curriculum than with M&I mechanics curriculum, thereby posing significant barriers to interpreting the meaning of FCI performance differences between traditional and M&I courses. We emphasize that these difficulties were not present in earlier work by some of us that used a concept inventory [the Brief Electricity and Magnetism Assessment (BEMA)] to evaluate comparatively traditional and M&I curricula for introductory electromagnetism.[14] In the earlier work, comparisons were made based on similar electromagnetism content present (in approximately equal measure) in both courses; moreover, the instrument used was carefully constructed to align with the minimal subset of content and goals in all courses.[20]

Notwithstanding its use to evaluate comparatively different curricula, data from the FCI might be used to adjust the content and goals of a given curriculum. For example, if faced with FCI performance similar to that reported here, M&I mechanics course instructors may make the (reasonable) decision that students should have more practice with qualitative

questions on topics covered by the FCI. In recent terms, we have made small modifications to the M&I curriculum by adding some homework problems and lab activities that are more aligned with the scope of the FCI. As a consequence, we have observed small improvements to the FCI scores of M&I students. We have not made a systematic study of which modifications to the M&I curriculum are most effective for improving student performance on the FCI.

The main purpose of an evaluation tool is to help answer the questions: Is the reform doing any good and, if so, is the good worth it? When the reform is curricular, concept inventories may be used to answer these questions when content and goals are shared by both curricula (with and without reform) with approximately equal intensity.[14] In the absence of this alignment, concept inventories might be used to give some insight into whether anything is "lost" with respect to overlapping content and goals. However, concept inventories (used to make comparisons) fail to measure what are perhaps the most interesting aspects of reform: the non-overlapping content and goals that are simply not present in courses without reform. In the case of M&I mechanics, examples of non-overlapping material include both new goals (e.g., relating macroscopic physics to microscopic models) (Ref. 16) and new content (e.g., computation).[18] There is a need to develop tools to help weigh the gains of a particular reform, so that instructors faced with multiple curricular choices can make informed decisions about which concepts, principles, and methods should be included or excluded, or emphasized or de-emphasized, during the finite time available to them to teach the course.

ACKNOWLEDGMENTS

The authors would like to thank the two anonymous reviewers for their many helpful comments. The authors would also like to thank Andrew Scherbakov and Robert Hume (Office of Minority Education and Development) for their efforts in collecting and organizing the demographic data. This work was supported by National Science Foundation's Division of Undergraduate Education (DUE0618519 and DUE0942076).

[a)]Electronic mail: marcos.caballero@colorado.edu; Present Address: Department of Physics, University of Colorado at Boulder, Boulder, CO 80309.
[b)]Electronic mail: michael.schatz@physics.gatech.edu

[1]P. M. Sadler and R. H. Tai, "Success in introductory college physics: The role of high school preparation," Sci. Educ. **85**(2), 111–136 (2001).
[2]U.S. Department of Education, National Center for Education Statistics, *Table 267: Degrees Conferred by Degree-Granting Institutions, by Control of Institution, Level of Degree, and Discipline Division: Selected years, 2005 through 2006* (Institute of Education Sciences, VA, 2007), <http://nces.ed.gov/programs/digest/d07/tables/dt07_267.asp>.
[3]P. J. Mulvey and S. Nicholson, *Physics Undergraduate Enrollments and Degrees* (AIP Statistical Research Center, MD, 2010), <http://www.aip.org/statistics/trends/reports/EDphysund07.pdf>.
[4]I. A. Halloun and D. Hestenes, "Common sense concepts about motion," Am. J. Phys. **53**(11), 1056–1065 (1985).
[5]E. F. Redish and R. N. Steinberg, "Teaching physics: Figuring out what works," Phys. Today **52**(1), 24–31 (1999).
[6]D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory," Phys. Teach. **30**(3), 141–158 (1992).
[7]R. Chabay and B. Sherwood, *Matter and Interactions I: Modern Mechanics*, 2nd ed. (John Wiley and Sons, Hoboken, NJ, 2007).
[8]L. McDermott, P. Shaffer, and the Physics Education Group at the University of Washington, *Tutorials in Introductory Physics*, 2nd. ed. (Pearson Education, Upper Saddle River, NJ, 2002).
[9]C. Wieman and K. Perkins, "Transforming physics education," Phys. Today **58**(11), 36–41 (2005).
[10]E. Mazur, *Peer Instruction: A User's Manual*, 1st ed. (Addison-Wesley, Boston, MA, 1997).
[11]E. S. Morote and D. E. Pritchard, "What course elements correlate with improvement on tests in introductory Newtonian mechanics?," Am. J. Phys. **77**(8), 746–753 (2009).
[12]E. Brewe, "Modeling theory applied: Modeling Instruction in introductory physics," Am. J. Phys. **76**(12), 1155–1160 (2008).
[13]M. Oliver-Hoyo and R. Beichner, "The SCALE-UP project," in *Teaching and Learning through Inquiry: A Guidebook for Institutions and Instructors*, edited by V.S. Lee (Stylus Publishing, Sterling, VA, 2004).
[14]M. A. Kohlmyer, M. D. Caballero, R. Catrambone, R. W. Chabay, L. Ding, M. P. Haugan, M. J. Marr, B. A. Sherwood, and M. F. Schatz, "A tale of two curricula: The performance of two thousand students in introductory electromagnetism," Phys. Rev. ST Phys. Educ. Res. **5**, 020105 (2009).
[15]R. D. Knight, *Physics for Scientists and Engineers: A Strategic Approach with Modern Physics w/Mastering Physics*, 1st ed. (Addison-Wesley, Boston, MA, 2004).
[16]R. Chabay and B. Sherwood, "Bringing atoms into first-year physics," Am. J. Phys. **67**(12), 1045–1050 (1999).
[17]R. Chabay and B. Sherwood, "Modern mechanics," Am. J. Phys. **72**(4), 439–445 (2004).
[18]R. Chabay and B. Sherwood, "Computational physics in the introductory calculus based course," Am. J. Phys. **76** (4&5), 307–313 (2008).
[19]M. A. Kohlmyer, "Student performance in computer modeling and problem solving in a modern introductory physics course," Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 2005.
[20]L. Ding, R. Chabay, B. Sherwood, and R. Beichner, "Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment," Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).
[21]D. Huffman and P. Heller, "What does the Force Concept Inventory actually measure?," Phys. Teach. **33**(3), 138–143 (1995).
[22]D. Hestenes and I. Halloun, "Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller," Phys. Teach. **33**(8), 502–506 (1995).
[23]P. Heller and D. Huffman, "Interpreting the force concept inventory: A reply to Hestenes and Halloun," Phys. Teach. **33**(8), 503–511 (1995).
[24]R. N. Steinberg and M. S. Sabella, "Performance on multiple-choice diagnostics and complementary exam problems," Phys. Teach. **35**(3), 150–155 (1997).
[25]N. S. Rebello and D. A. Zollman, "The effect of distracters on student performance on the force concept inventory," Am. J. Phys. **72**(1), 116–125 (2004).
[26]R. R. Hake, "Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66**(1), 64–74 (1998).
[27]X. H. Zhou and P. Dinh, "Nonparametric confidence intervals for one-and two-sample problems," Biostatistics **6**(2), 187–200 (2005).
[28]W. J. Conover, *Practical Nonparametric Statistics*, 2nd. ed. (John Wiley and Sons, Hoboken, NJ, 1999).
[29]P. Sprent, *Applied Nonparametric Statistical Methods*, 2nd. ed. (Chapman and Hall, London, England, 1993).
[30]L. Ding, N. W. Reay, A. Lee, and L. Bao, "Effects of testing conditions on conceptual survey results," Phys. Rev. ST Phys. Educ. Res. **4**, 010112 (2008).
[31]C. H. Crouch and E. Mazur, "Peer instruction: Ten years of experience and results," Am. J. Phys. **69**(9), 970–977 (2001).
[32]S. J. Pollock, S. V. Chasteen, M. Dubson, and K. K. Perkins, "The use of concept tests and peer instruction in upper-division physics," in *AIP Conference Proceedings*, edited by M. Sabella, C. Singh, and S. Rebello (AIP Press, NY, 2010), Vol. 1289, p. 261.