# Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses*

Richard R. Hake[a)]
*Department of Physics, Indiana University, Bloomington, Indiana 47405*

A survey of pre/post test data using the Halloun-Hestenes Mechanics Diagnostic test or more recent Force Concept Inventory is reported for 62 introductory physics courses enrolling a total number of students N = 6542. A consistent analysis over diverse student populations in high schools, colleges, and universities is obtained if a rough measure of the average effectiveness of a course in promoting conceptual understanding is taken to be the average normalized gain <g>. The latter is defined as the ratio of the actual average gain (%<post> – %<pre>) to the maximum possible average gain (100 – %<pre>). Fourteen "traditional" (T) courses (N = 2084) which made little or no use of interactive-engagement (IE) methods achieved an average gain $\langle g \rangle_{T\text{-}ave} = 0.23 \pm 0.04$ (std dev). In sharp contrast, forty-eight courses (N = 4458) which made substantial use of IE methods achieved an average gain $\langle g \rangle_{IE\text{-}ave} = 0.48 \pm 0.14$ (std dev), almost two standard deviations of $\langle g \rangle_{IE\text{-}ave}$ above that of the traditional courses. Results for 30 (N = 3259) of the above 62 courses on the problem-solving Mechanics Baseline test of Hestenes-Wells imply that IE strategies enhance problem-solving ability. The conceptual and problem-solving test results strongly suggest that the classroom use of IE methods can increase mechanics-course effectiveness well beyond that obtained in traditional practice.

## I. INTRODUCTION

There has been considerable recent effort to improve introductory physics courses, especially after 1985 when Halloun and Hestenes[1] published a careful study using massive pre- and post-course testing of students in both calculus and non-calculus-based introductory physics courses at Arizona State University. Their conclusions were: (1) "....the student's initial qualitative, common-sense beliefs about motion and....(its).... causes have a large effect on performance in physics, but conventional instruction induces only a small change in those beliefs." (2) "Considering the wide differences in the teaching styles of the four professors....(involved in the study)....the basic knowledge gain under conventional instruction is essentially independent of the professor." These outcomes were consistent with earlier findings of many researchers in physics education (see refs. 1 - 8 and citations therein) which suggested that traditional passive-student introductory physics courses, even those delivered by the most talented and popular instructors, imparted little conceptual understanding of Newtonian mechanics.

To what extent has the recent effort to improve introductory physics courses succeeded? In this article I report a survey of all quantitative pre/post test results known to me (in time to be included in this report) which use the original Halloun-Hestenes Mechanics Diagnostic test (MD),[1a] the more recent Force Concept Inventory (FCI),[9a,b] and the problem-solving Mechanics

* Accepted for publication in the *American Journal of Physics.* Comments and criticisms will be welcomed at R.R. Hake, 24245 Hatteras St., Woodland Hills, CA, USA 91367, <hake@ix.netcom.com>.

Baseline (MB)[10] test. Both the MD and FCI were designed to be tests of students' conceptual understanding of Newtonian mechanics. One of their outstanding virtues is that the questions probe for conceptual understanding of basic concepts of Newtonian mechanics in a way that is understandable to the novice who has never taken a physics course, while at the same time rigorous enough for the initiate.

Most physicists would probably agree that a low score on the FCI/MD test indicates a lack of understanding of the basic concepts of mechanics. However, there have been recent con[11] and pro[12] arguments as to whether a high FCI score indicates the attainment of a unified force concept. Nevertheless, even the detractors have conceded that "the FCI is one of the most reliable and useful physics tests currently available for introductory physics teachers"[11a] and that the FCI is "the best test currently available .... to evaluate the effectiveness of instruction in introductory physics courses."[11b] While waiting for the fulfillment of calls for the development of better tests[11] or better analyses of existing tests,[12] the present survey of published[1a,8a,9a,13,14] and unpublished[15a,b] classroom results may assist a much needed further improvement in introductory mechanics instruction in the light of practical experience.

## II. SURVEY METHOD AND OBJECTIVE

Starting in 1992, I requested that pre/post FCI test data and posttest MB data be sent to me in talks at numerous colloquia and meetings and in e-mail postings on the PHYS-L and PhysLrnR nets.[16] This mode of data solicitation tends to pre-select results which are biased in favor of outstanding courses which show relatively high gains on the FCI. When relatively low gains are achieved (as they often are) they are sometimes mentioned informally, but they are usually neither published nor communicated except by those who (a) wish to use the results from a "traditional" course at their institution as a baseline for their own data, or (b) possess unusual scientific objectivity and detachment. Fortunately, several in the latter category contributed data to the present survey for courses in which interactive engagement methods were used but relatively low gains were achieved. Some suggestions (Sec. VII) for increasing course effectiveness have been gleaned from those cases.[17]

Some may think that the present survey presents a negatively biased sampling of traditional courses, an attitude which has been known to change after perusal of local FCI test results.[18] It should be emphasized that all traditional-course pre/post test data known to me in time to be included in this report are displayed in Fig. 1. More such data undoubtedly exists but goes unreported because the gains are so embarrassingly minimal.

For survey classification and analysis purposes I define:

(a) "Interactive Engagement" (IE) methods as those *designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors,* all as judged by their literature descriptions;

(b) "Traditional" (T) courses as those reported by instructors to *make little or no use of IE methods, relying primarily on passive-student lectures, recipe labs, and algorithmic-problem exams;*

(c) "Interactive Engagement" (IE) courses as those reported by instructors to *make substantial use of IE methods*;

(d) average normalized gain <g> for a course as the ratio of the actual average gain <G> to the maximum possible average gain, i.e.,

$$\langle g \rangle \equiv \%\langle G \rangle \; / \; \%\langle G \rangle_{max}$$

$$= (\%\langle S_f \rangle - \%\langle S_i \rangle) / (100 - \%\langle S_i \rangle), \quad \text{.....(1)}$$

where $\langle S_f \rangle$ and $\langle S_i \rangle$ are the final (post) and initial (pre) class averages;

(e) "High-g" courses as those with $(\langle g \rangle) \geq 0.7$;

(f) "Medium-g" courses as those with $0.7 > (\langle g \rangle) \geq 0.3$;

(g) "Low-g" courses as those with $(\langle g \rangle) < 0.3$.

The present survey covers 62 introductory courses enrolling a total of 6542 students using the conceptual MD or FCI exams, and (where available) the problem-solving Mechanics Baseline (MB) test. Survey results for the conceptual and problem-solving exams are presented below in the form of graphs. In a companion paper,[17a] intended to assist instructors in selecting and implementing proven IE methods, I tabulate, discuss, and reference the particular methods and materials that were employed in each of the 62 survey courses. Also tabulated in ref. 17a are data for each course: instructor's name and institution, number of students enrolled, pre/post test scores, standard deviations where available, and normalized gains. Survey information was obtained from published accounts or private communications. The latter usually included instructor responses to a survey questionnaire[15c] which asked for information on the pre/post testing method; statistical results; institution; type of students; activities of the students; and the instructor's educational experience, outlook, beliefs, orientation, resources, and teaching methods.

As in any scientific investigation, bias in the detector can be put to good advantage if appropriate research objectives are established. We do *not* attempt to access the *average* effectiveness of introductory mechanics courses. Instead we seek to answer a question of considerable practical interest to physics teachers: *Can the classroom use of IE methods increase the effectiveness of introductory mechanics courses well beyond that attained by traditional methods*?

## III. CONCEPTUAL TEST RESULTS

### A. Gain vs Pretest Graph - All Data

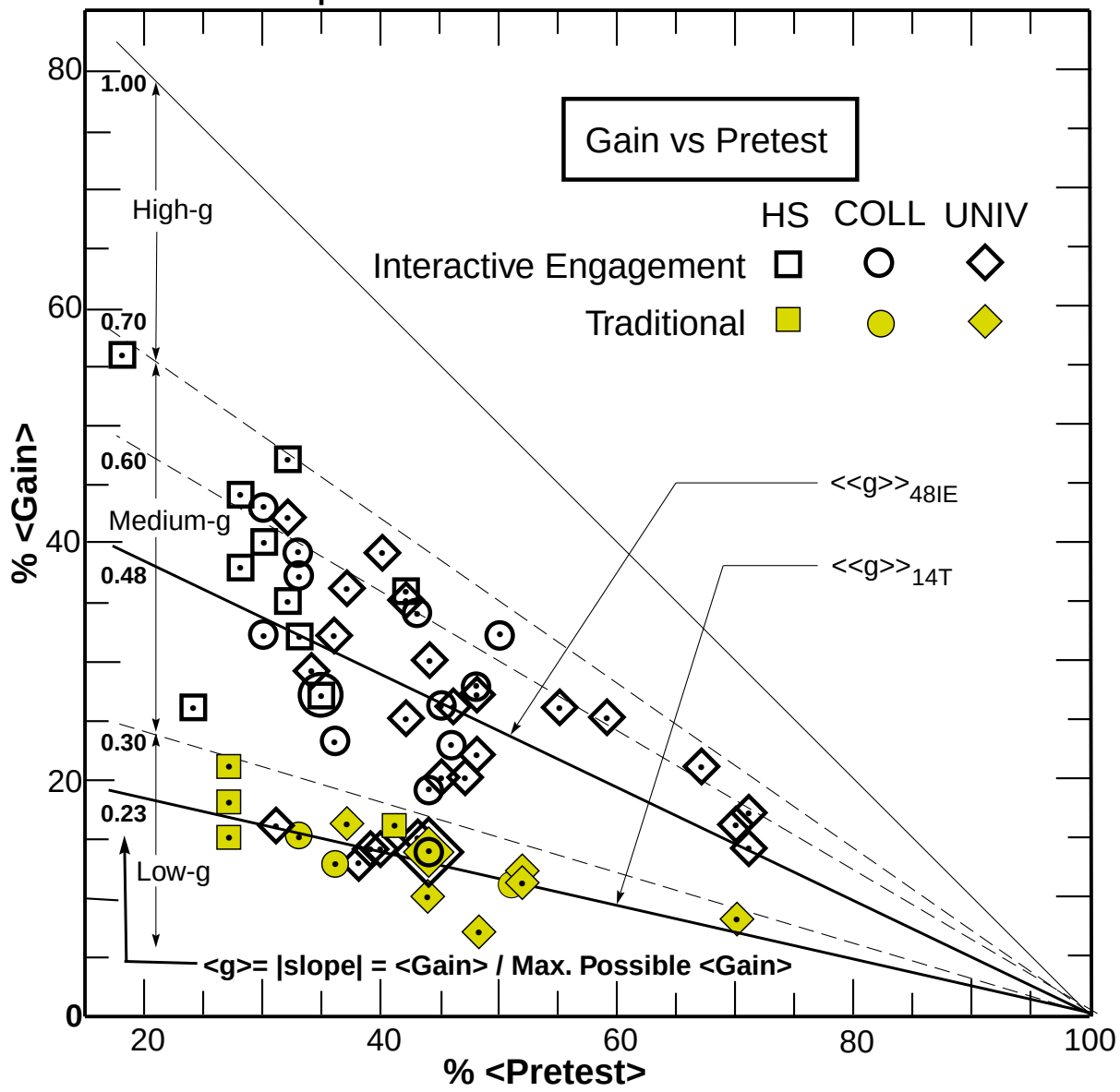


Fig. 1. %<Gain> vs %<Pretest> score on the conceptual Mechanics Diagnostic (MD) or Force Concept Inventory (FCI) tests for 62 courses enrolling a total N = 6542 students: 14 traditional (T) courses (N = 2084) which made little or no use of interactive engagement (IE) methods, and 48 IE courses (N = 4458) which made considerable use of IE methods. Slope lines for the average of the 14 T courses $\langle\langle g \rangle\rangle_{14T}$ and 48 IE courses $\langle\langle g \rangle\rangle_{48IE}$ are shown, as explained in the text.

To increase the statistical reliability (Sec. IV) of averages *over courses*, only those with enrollments $N \geq 20$ are plotted in Fig. 1, although in some cases of fairly homogeneous instruction and student population (AZ-AP, AZ-Reg, PL92-C, TO, TO-C) courses or sections with less than 20 students were included in a number-of-student-weighted average. Course codes such as "AZ-AP" with corresponding enrollments and scores are tabulated and referenced in ref.

17a. In assessing the FCI, MD, and MB scores it should be kept in mind that the random guessing score for each of these five-alternative multiple-choice tests is 20%. However, completely non-Newtonian thinkers (if they can at the same time read and comprehend the questions) may tend to score *below* the random guessing level because of the very powerful interview-generated distractors.[1a,12a]

It should be noted that for any particular course point ($\langle G' \rangle$, $\langle S_i' \rangle$) on the $\langle G \rangle$ vs $\langle S_i \rangle$ plot of Fig. 1, the absolute value of the slope of a line connecting ($\langle G' \rangle$, $\langle S_i' \rangle$) with the point ($\langle G \rangle = 0$, $\langle S_i \rangle = 100$) is just the gain parameter $\langle g' \rangle$ for that particular course. The regularities for courses with a wide range of average pretest scores [$18 \leq (\langle S_i \rangle) \leq 71$] and with diverse student populations in high schools, colleges, and universities are noteworthy:

(a) *All* points for the 14 T courses (N = 2084) fall in the Low–g region. The data[17a] yield

$$\langle\langle g \rangle\rangle_{14T} = 0.23 \pm 0.04\text{sd} \quad \text{..... (2a)}$$

Here and below, double carets "$\langle\langle X \rangle\rangle_{NP}$" indicate an average of averages, i.e., an average of $\langle X \rangle$ over N courses of type P, and sd ≡ standard deviation [*not* to be confused with random or systematic experimental error (Sec. V)].

(b) Eighty-five percent (41 courses, N = 3741) of the 48 IE courses fall in the Medium-g region and 15% (7 courses, N = 717) in the Low-g region. Overall, the data[17a] yield

$$\langle\langle g \rangle\rangle_{48IE} = 0.48 \pm 0.14\text{sd}. \quad \text{..... (2b)}$$

The slope lines $\langle\langle g \rangle\rangle$ of Eq. (2a,b) are shown in Fig. 1.

(c) No course points lie in the "High-g" region.

I infer from features a, b, c that a consistent analysis over diverse student populations with widely varying initial knowledge states, as gauged by $\langle S_i \rangle$, 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 $\langle S_i \rangle$ for the 62 survey courses is a very low +0.02. In contrast, the average posttest score $\langle S_f \rangle$ and the average gain $\langle G \rangle$ are less suitable for comparing course effectiveness over diverse groups since their correlations with $\langle S_i \rangle$ are, respectively, +0.55 and –0.49. It should be noted that a positive correlation of $\langle S_f \rangle$ with $\langle S_i \rangle$ would be expected in the absence of instruction.

Assuming, then, that $\langle g \rangle$ is a valid measure of course effectiveness in promoting conceptual understanding, it appears that the present interactive engagement courses are, on average, more than twice as effective in building basic concepts as traditional courses since $\langle\langle g \rangle\rangle_{IE} = 2.1 \langle\langle g \rangle\rangle_{T}$. The difference

$$\langle\langle g \rangle\rangle_{48IE} - \langle\langle g \rangle\rangle_{14T} = 0.25 \quad \text{.....(2c)}$$

is 1.8 standard deviations of $\langle\langle g \rangle\rangle_{48IE}$ and 6.2 standard deviations of $\langle\langle g \rangle\rangle_{14T}$, reminiscent of that seen in comparing instruction delivered to students in large groups with one-on-one instruction.[19]

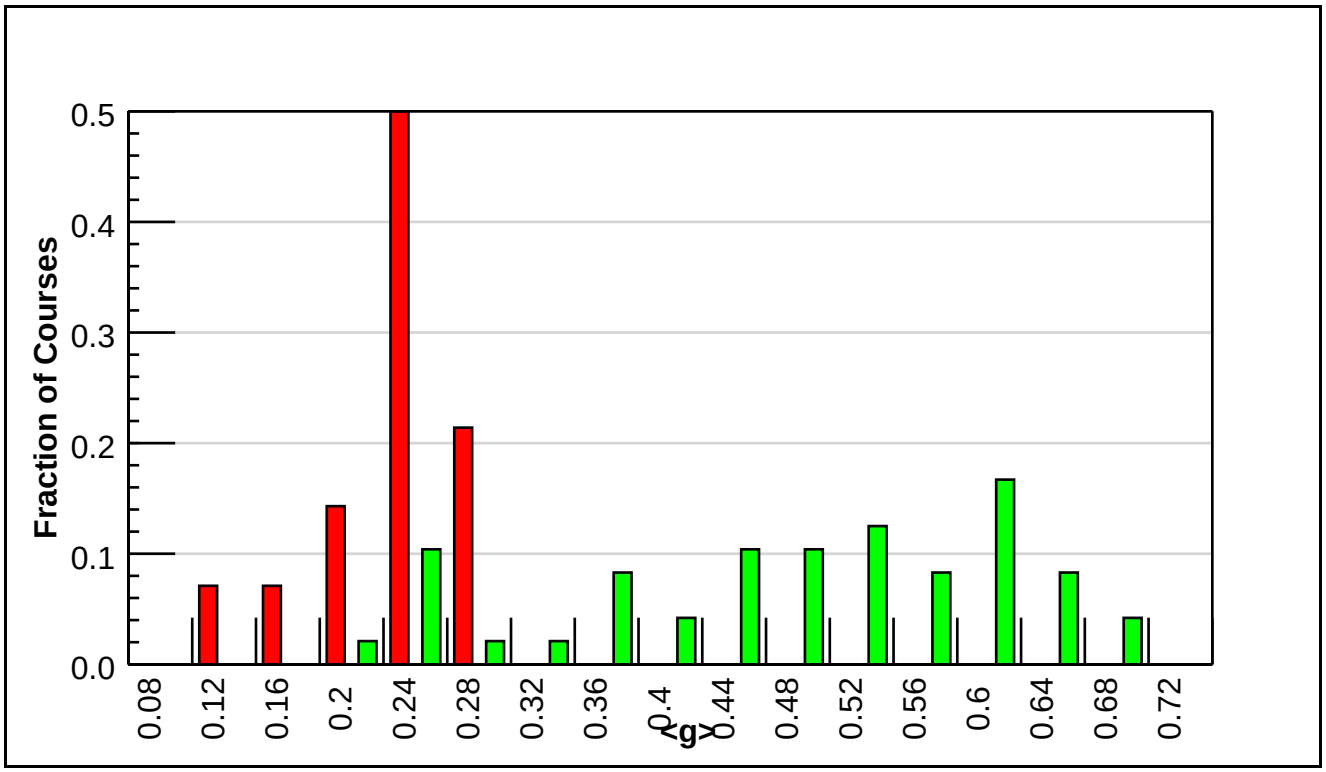


Fig. 2. Histogram of the average normalized gain <g>: dark (red) bars show the *fraction* of 14 traditional courses (N = 2084), and light (green) bars show the *fraction* of 48 interactive engagement courses (N = 4458), both within bins of width $\delta\langle g\rangle = 0.04$ centered on the <g> values shown.

Figure 2 shows the <g>-distribution for *traditional* (T) and *interactive engagement* (IE) courses plotted in Fig. 1. Both distributions deviate from the symmetric Gaussian shape, but this does not invalidate characterization of the spread in the data by the standard deviation.

The widths of the <g> distributions are evidently related to (a) statistical fluctuations in <g> associated with widths of the pre- and posttest score distributions as gauged by their standard deviations, plus (b) course-to-course variations in the "systematic errors," plus (c) course-to-course variations in the effectiveness of the pedagogy and/or implementation. I use the term "systematic errors" to mean that *for a single course* the errors would affect test scores in a systematic way, even though such errors might affect different courses in a more-or-less random way. Statistical fluctuations and systematic errors in <g> are discussed below in Sec.V. Case studies[17a] of the IE courses in the low-end bump of the IE distribution strongly suggest that this bump is related to "c" in that various implementation problems are apparent: e.g., insufficient training of instructors new to IE methods, failure to communicate to students the nature of science and learning, lack of grade incentives for taking IE activities seriously, a paucity of exam questions which probe the degree of conceptual understanding induced by the IE methods, and use of IE methods in only isolated components of a course.

**B. Gain vs Pretest Graphs for High Schools, Colleges, and Universities**

Figures 3a,b,c show separate G vs $S_i$ plots for the 14 high school (N = 1113), 16 college (N = 597), and 32 university courses (N = 4832). Although the *enrollment N-weighted* average pretest scores increase with level[20] [$\langle S_i \rangle_{HS}$ = 28%, $\langle S_i \rangle_C$ = 39%, $\langle S_i \rangle_U$ = 48% (44% if the atypically high Harvard scores are omitted)], in other respects these three plots are all very similar to the plot of Fig. 1 for all courses. For high schools, colleges, and universities (a) T courses achieve low gains close to the average $\langle\langle g \rangle\rangle_{T14}$ = 0.23; (b) IE courses are about equally effective: $\langle\langle g \rangle\rangle_{10IE(HS)}$ = 0.55 ± 0.11sd, $\langle\langle g \rangle\rangle_{13IE(C)}$ = 0.48 ± 0.12sd, and $\langle\langle g \rangle\rangle_{25IE(U)}$ = 0.45 ± 0.15sd (0.53 ± 0.09sd if the averaging omits the 6 atypical Low-g university courses).

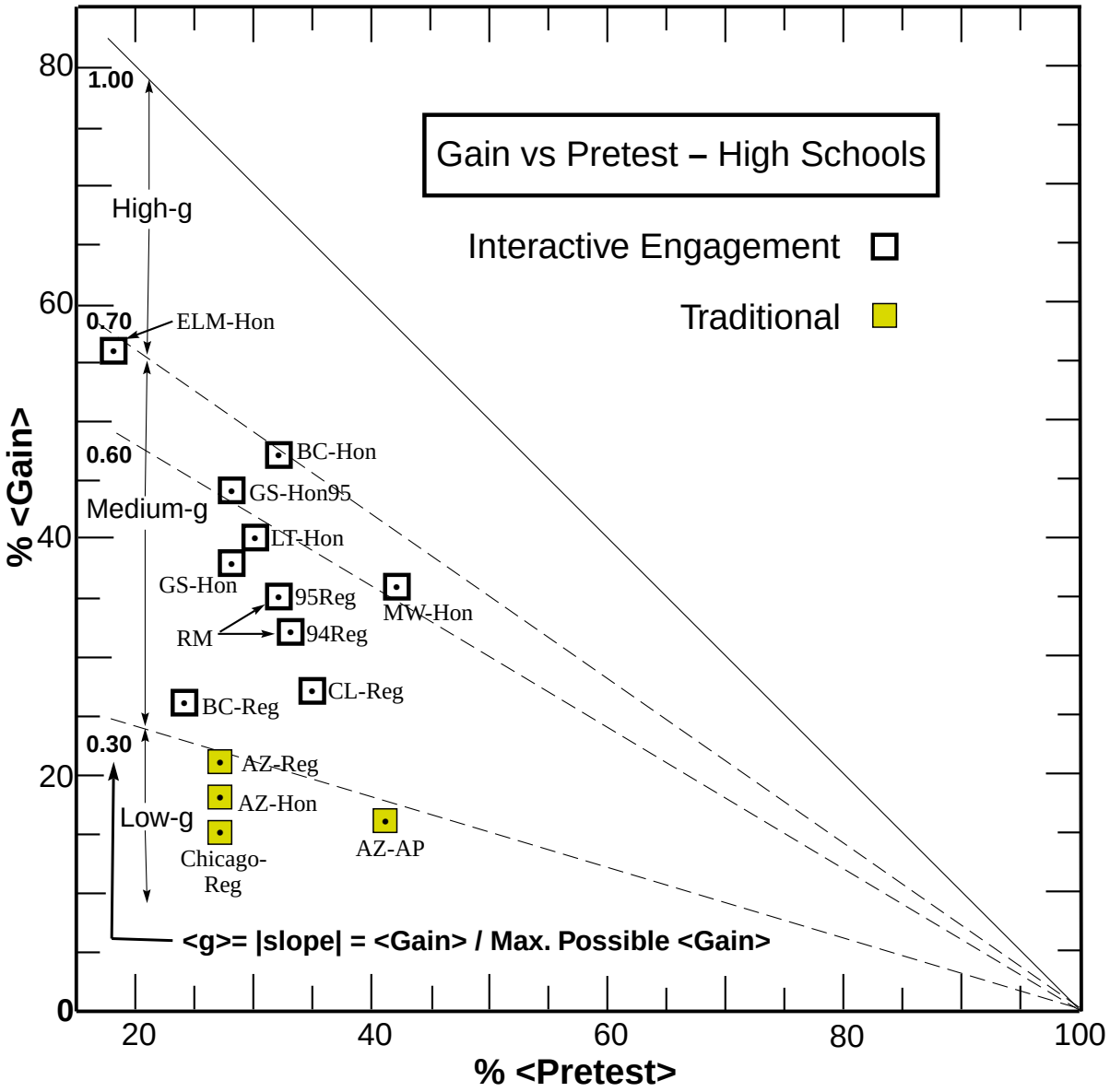


Fig. 3a. %<Gain> vs %<Pretest> score on the conceptual Mechanics Diagnostic (MD) or Force Concept Inventory (FCI) tests for 14 *high-school* courses enrolling a total of N = 1113 students. In this and subsequent figures, course codes, enrollments, and scores are tabulated and referenced in ref. 17a.

Fig. 3a shows that, for high schools, higher g's are obtained for honors than for regular courses, consistent with the observations of Hestenes *et al.*[9a] The difference between these two groups is perceived differently by different instructors and may be school dependent: "the main difference is attitude"[9a]; "they differ in their ability to use quantitative representations of data to draw conceptual generalizations....motivation is.... only part of the difference"[21]; "both sets... (are)... highly motivated....the major differences....(are).... their algebraic skills, the degree of confidence in themselves, their ability to pay attention to detail, and their overall ability."[22] Motivational problems can be especially severe for students in IE courses who dislike any departure from the traditional methods to which they have become accustomed and under which their grades, if not their understanding, may have flourished.[23-26]

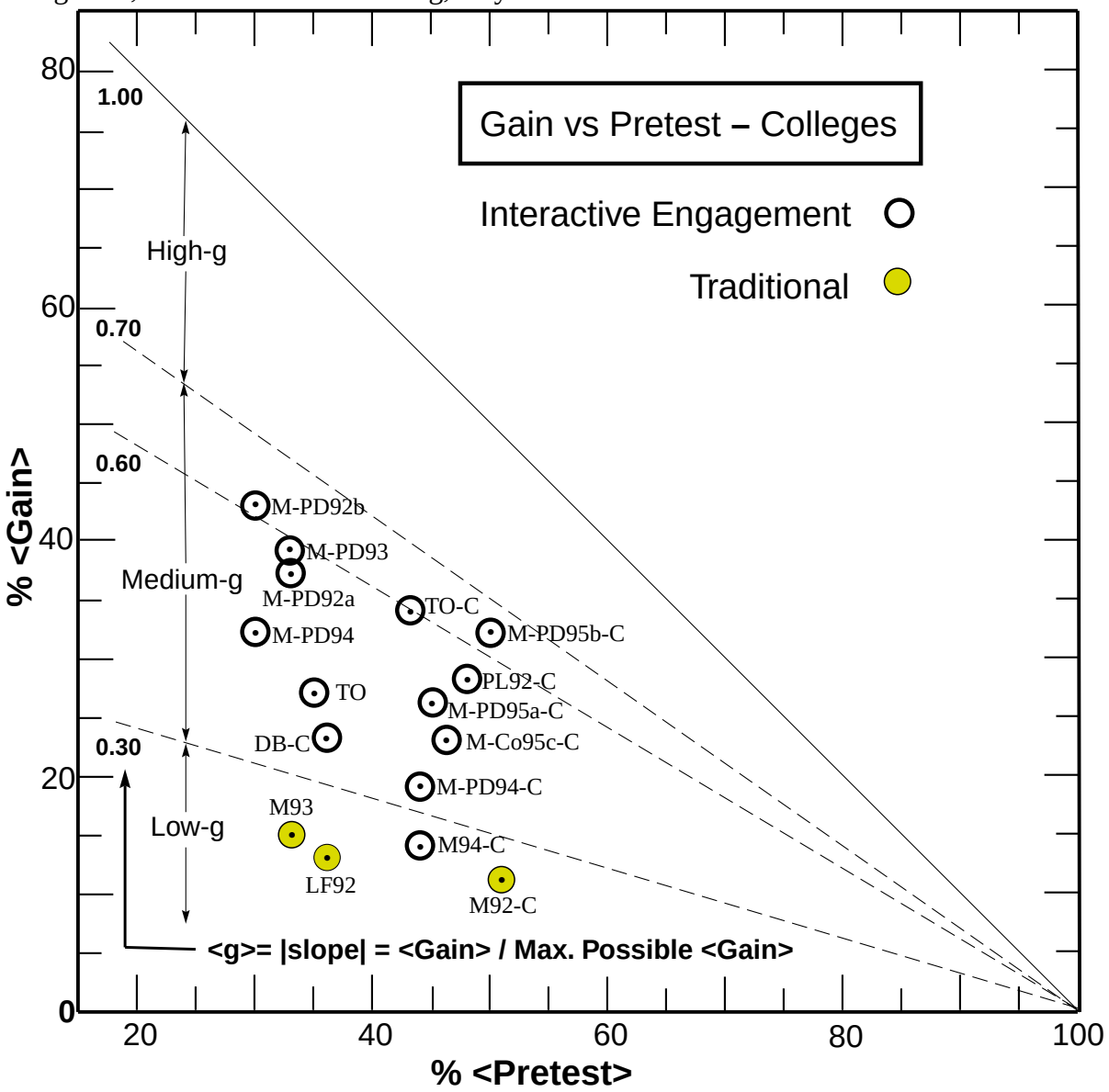


Fig. 3b. %<Gain> vs %<Pretest> score on the conceptual MD or FCI tests for 16 *college* courses enrolling a total of N = 597 students. The course code "-C" indicates a calculus-based course.

Enrollments for the college courses of Fig. 3b are in the 20 - 61 range so that statistical fluctuations associated with "random errors" (Sec. V) could be relatively important. However the variations in <g> for the eleven Monroe Community College courses (M) have been explained[17a] by Paul D'Alessandris[27] as due to differences in the students or in the instruction: e.g., "With regard to the....<g> differences in.... the two sections of calculus-based physics in 1995, M-PD95b-C....<g> = 0.64....was a night course and M-PD95a-C....<g> = 0.47.... was a day course. The difference in the student populations between night and day school is the difference between night and day. The night students average about 7-10 years older, much more mature and dedicated, possibly because they are all paying their own way through school. The actual instructional materials and method were the same for both groups. The instructional materials do change semester by semester (I hope for the better).... M-PD94-C had <g> = 0.34 (this was the first time I used my materials in a calculus-based class.) M-PD95a-C had <g> = 0.47, and in the Fall of 1995....not included in this survey because N = 15..... I had a <g> of 0.63. This change is, hopefully, not a random fluctuation but due to the changes in the workbook. All these were day courses." Such tracking of <g> with changes in IE method or implementation, also observed at Indiana University[17a] enhances confidence in the use of <g> as a gauge of course effectiveness in building basic concepts.

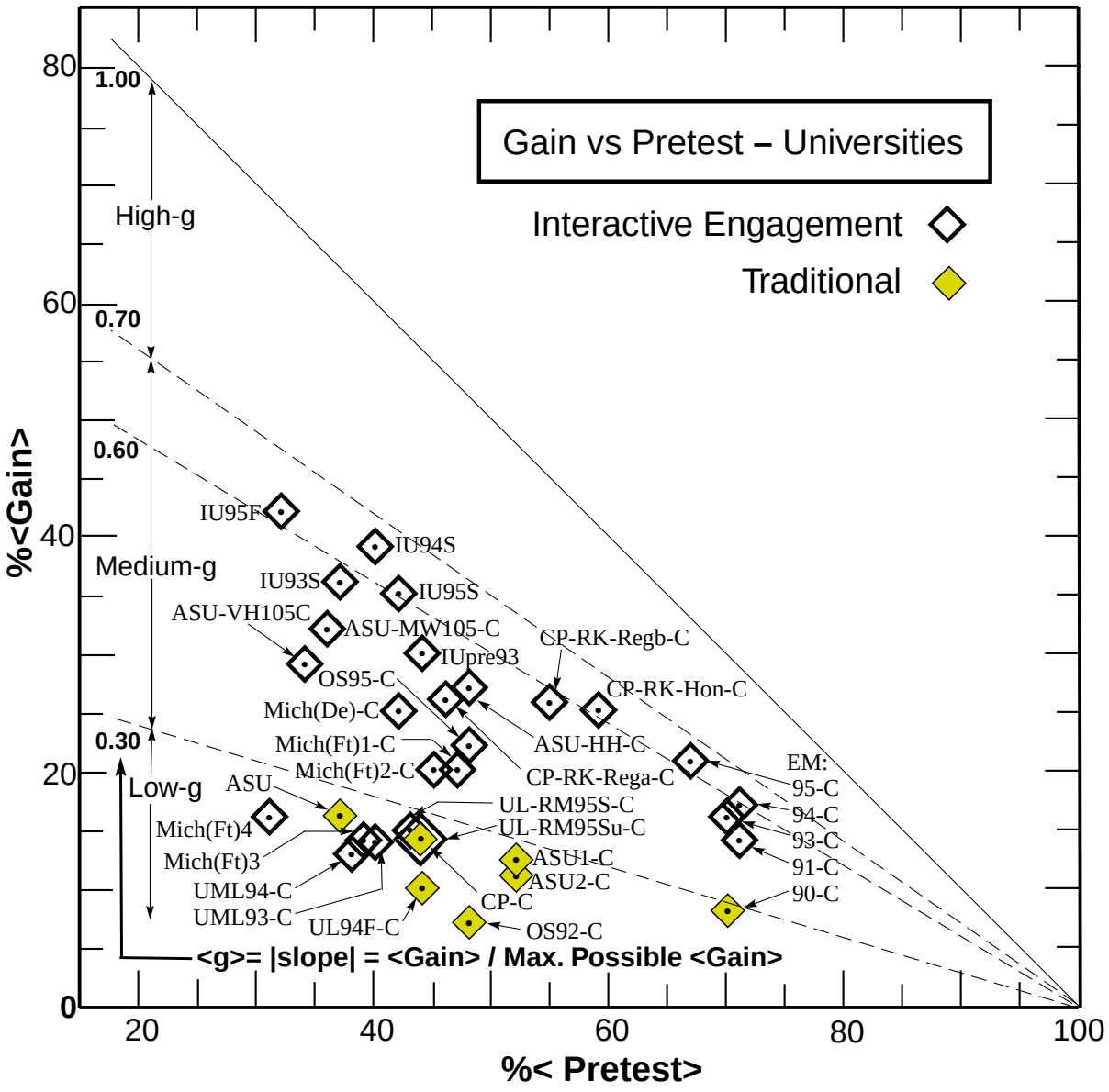


Fig. 3c. %<Gain> vs %<Pretest> score on the conceptual MD or FCI tests for 32 university courses enrolling a total N = 4832 students. The course code "-C" indicates a calculus-based course.

For university courses (Fig. 3c) six of the IE courses are in the Low-g region – as previously indicated, detailed case studies[17a] strongly suggest that implementation problems are responsible. Ten[17a] of the IE courses in the Medium-g region have enrollments over 100 and four have enrollments over 200 – OS95-C: 279; EM94-C: 216; IU95S: 209; IU95F: 388. All the N > 200 courses[28a;29a;30c,d] attempt to bring IE methods to the masses in cost-effective ways by means of (a) collaborative peer instruction[31,32] and (b) employment of undergraduate students to augment the instructional staff (Sec. VII).

The work at Ohio State is part of an ongoing and concerted *departmental effort*, starting in 1993, and actively involving about 30% of the faculty.[28a] The long-range goal is to induce a badly needed (see the point for OS92-C in Fig. 3c) systemic improvement in the effectiveness of all the introductory courses. The largest-enrollment introductory physics course at Ohio State, of concern here, is designed for engineering students. In this course there is an unusually heavy emphasis on "using symbolic language with understanding to solve complex problems." In addition to "a" and "b," above, use is made of: (1) Overview Case Studies (OCS),[28b] Active Learning Problem Sets (ALPS)[28b] with context-rich problems,[32] and interactive simulations with worksheets;[28a] all of these in interactive "lectures" (called "Large Room Meetings"); (2) cooperative group problem-solving of context-rich problems and multiple-representation exercises in "recitations" (called "Small Room Meetings"); (3) an inquiry approach with qualitative questions and experiment problems[28c] in the labs.

Harvard adds Concept Tests,[29b,c] a very complete course Web page,[29c] and computer communication between and among students and instructors,[29c,33] to "a" and "b."

Indiana University adds to to "a" and "b": SDI labs[8,13,34,35]; Concept Tests[29,36]; cooperative group problem-solving in "recitations"[32,37,38]; computer communication between and among students and instructors[39]; Minute Papers[37,40]; team teaching[30c,d]; a mid-course *diagnostic* student evaluation over all aspects and components of the course[13,41]; an academic background questionnaire[8a,13] which allows instructors to become personally familiar with the aspirations and preparation of each incoming student; a "Physics Forum" staffed by faculty and graduate students for 5-8 hours/day where introductory students can find help at any time[38]; color coding[8a,13,34] of displacement, velocity, acceleration, and force vectors in *all* components of the course; and the use of grading acronyms[34e] to increase the efficiency of homework grading (e.g., NDC ≡ Not Dimensionally Correct).

## IV. MECHANICS BASELINE TEST RESULTS

The Mechanics Baseline test is designed to measure more quantitative aspects of student understanding than the FCI. It is usually given only as a posttest. Figure 4 shows a plot of the average percentage score on the problem-solving Mechanics Baseline (MB) posttest vs the average percentage score on the FCI posttest for all the available data.[17a] The solid line is a least-squares fit to the data points. The two scores show an extremely strong positive correlation with coefficient $r = +0.91$. Such a relationship is not unreasonable because the MB test (unlike most traditional algorithmic-problem physics exams) requires conceptual understanding in addition to some mathematical skill and critical thinking. Thus the MB test is more difficult for the average student, as is also indicated by the fact that MB averages tend to be about 15% below FCI averages, i.e., the least-squares-fit line is nearly parallel to the diagonal (%MB = %FCI) and about 15% points below it.[42]

It is sometimes objected that the problems on the MB test do not sufficiently probe more advanced abilities such as those required for problems known as: "context rich"[32]; "experiment"[28c]; "goal-less" [27b]; "out-of-lab"[34c]; or Fermi. On the other hand, some instructors object that neither the MB problems nor those indicated above are "real" problems because they are somewhat different from "Halliday-Resnick problems." Considering the differences in outlook, it may be some time before a more widely accepted problem-solving test becomes available.

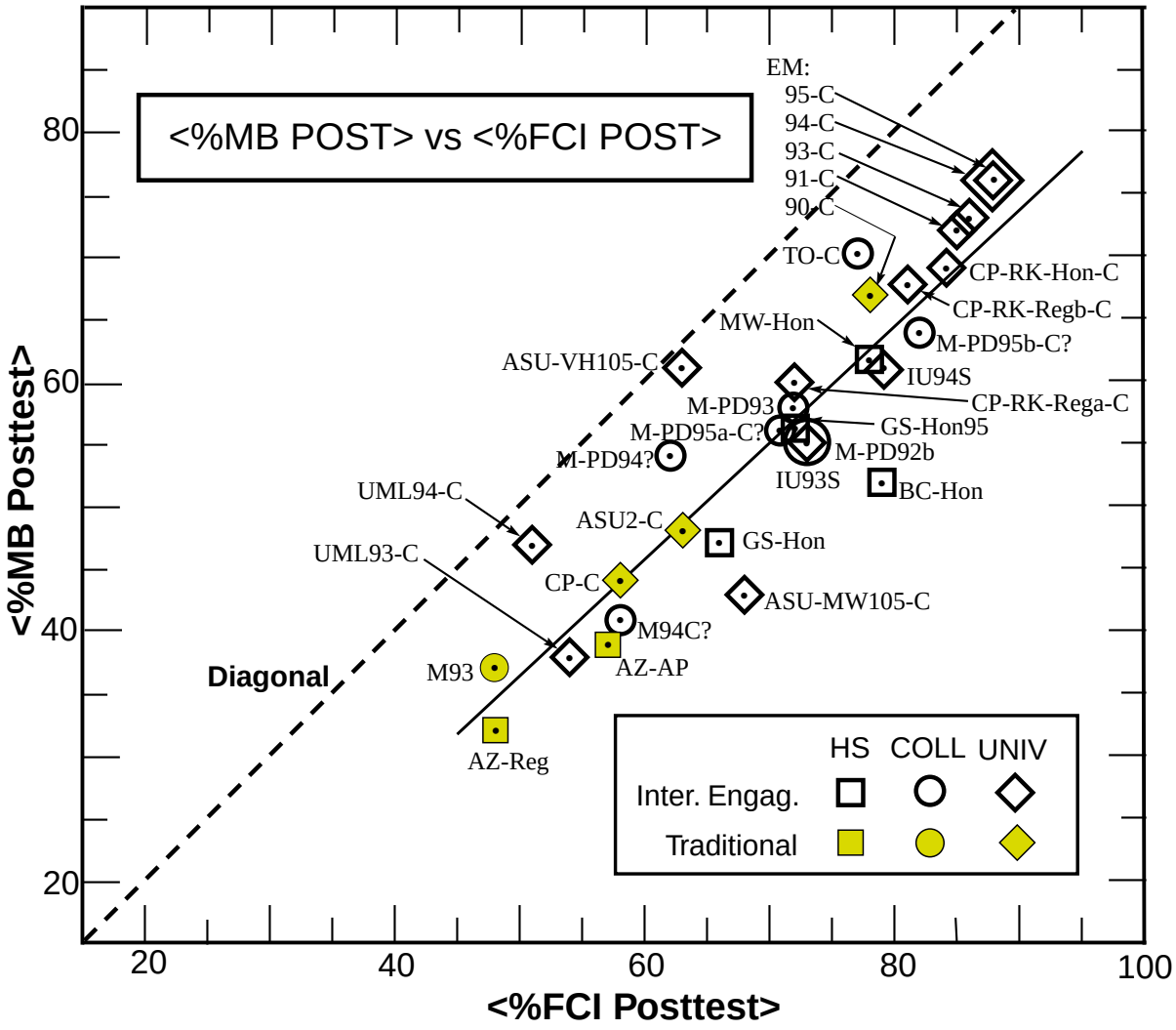


Fig. 4. Average posttest scores on the problem-solving Mechanics Baseline (MB) test vs those on the conceptual FCI test for all courses of this survey for which data are available: thirty courses (high school, college, and university) which enroll a total N = 3259 students (ref. 17a). The solid line is a least-squares fit to the data points. The dashed line is the diagonal representing equal scores on the MB and FCI tests. Courses at Monroe Community College (M) with a "?" designation had non-matching $N_{MB} > N_{FCI}$ because a few students who took the MB did not also take the FCI pretest, as indicated in ref. 17a. If these "?" points are excluded from the analyses, then the correlation coefficient "r" changes by less than 0.1% and the change in the position of the least-squares-fit line is almost imperceptible on the scale of this figure.

Figure 4 shows that IE courses generally show both higher FCI averages and higher MB averages than traditional courses, especially when the comparison is made for courses with similar student populations, e.g., Cal Poly [CP-C vs (CP-RK-Rega-C, CP-RK-Regb-C, and CP-RK-Hon-C)]; Harvard (EM90-C vs EM91,93,94,95-C); Monroe Community College (MCC) [M93 vs other M-prefix courses]; Arizona high schools [(AZ-Reg & AZ-AP) vs MW-Hon]. Thus it would appear that problem-solving capability is actually *enhanced* (not sacrificed as some would believe) when concepts are emphasized. This is consistent with the observations of Mazur[29b] and with the results of Thacker *et al.*,[43] showing that, at Ohio State, elementary-education majors taking an inquiry-based course did better than students enrolled in a conventional physics courses for engineers on both a synthesis problem and an analysis problem.

## V. ERRORS IN THE NORMALIZED GAIN

### A. Statistical Fluctuations ("Random Errors")

The widths of the distributions of pre- and posttest scores as characterized by their standard deviations (7 to 21% of the total number of questions on the exam[17a]) are quite large. In most cases these widths are not the result of experimental error but primarily reflect the varying characteristics of the students. If a multiplicity of understandings, abilities, skills, and attitudes affect test performance and these vary randomly among the students then a near Gaussian distribution would be expected for high N. Redish[44] calls this "the individuality or 'linewidth' principle." The large linewidths create "random error" uncertainties in the pre- and posttest averages and therefore statistical fluctuations ("random errors") $\Delta\langle g\rangle$ in the average normalized gains $\langle g\rangle$. I have calculated $\Delta\langle g\rangle$'s in the conventional manner[45,46] for the 33 survey courses for which deviations are available.[17a] For this subset :

$$\langle\langle g\rangle\rangle_{T9} = 0.24 \pm 0.03\text{sd}, \quad \text{.....(3a)}$$

$$\langle\langle g\rangle\rangle_{IE24} = 0.50 \pm 0.12\text{sd}, \quad \text{.....(3b)}$$

similar to the averages and standard deviations for all the data as indicated in Eq. (2a,b). The random error averages $\langle(\Delta\langle g\rangle)\rangle$ for the subset are

$$\langle(\Delta\langle g\rangle)\rangle_{T9} = 0.04 \pm 0.02\text{sd}, \quad \text{.....(4a)}$$

$$\langle(\Delta\langle g\rangle)\rangle_{IE24} = 0.04 \pm 0.02\text{sd} . \quad \text{.....(4b)}$$

According to the usual interpretation,[45] if only random errors are present then the standard deviation for an average of averages, Eq. (3), should be about the same as the uncertainty in any one average, Eq. (4). (For a numerical example see ref. 45b.) This would suggest that, for the subset, the spread (sd = 0.03) in the $\langle g\rangle_T$ distribution can be accounted for primarily by random-errors [$\langle(\Delta\langle g\rangle)\rangle_{T9} = 0.04$], while the spread (sd = 0.12) in the $\langle g\rangle_{IE}$ distribution is due to random errors [$\langle(\Delta\langle g\rangle)\rangle_{IE24} = 0.04$] *plus other factors*: course-to-course variation in the systematic error, and course-to-course variation in the effectiveness of the pedagogy and/or implementation.

### B. Systematic Error

Aside from the previously mentioned controversy[11,12] over the interpretation of a high FCI score, criticism of FCI testing sometimes involves perceived difficulties such as (1) question ambiguities and isolated false positives (right answers for the wrong reasons); and uncontrolled variables in the testing conditions such as (2) teaching to the test and test-question leakage, (3) the fraction of course time spent on mechanics, (4) post and pretest motivation of students, and (5) the Hawthorne/John Henry effects.[47]

For both IE and T courses, the influence of errors "2'' through "5" would be expected to vary from course to course in a more or less random manner, resulting in a systematic-error "noise" in gain vs pretest plots containing data from many courses. Although the magnitude of this noise is difficult to estimate, it contributes to the width of the $\langle g\rangle$ distributions specified in Eq. (2). The analysis of random errors above suggests that the systematic-error noise and the course-to-course variations in the effectiveness of the pedagogy and/or implementation contribute more importantly to the width of the $\langle g\rangle_{IE}$ distribution than to the width of the $\langle g\rangle_T$ distribution.

It is, of course, possible that the systematic errors, even though varying from course-to-course, could, on average, positively bias the IE gains so as to increase the difference $\langle\langle g\rangle\rangle_{IE48} - \langle\langle g\rangle\rangle_{T14}$. I consider below each of the above-indicated systematic errors.

***1. Question Ambiguities and Isolated False Positives***

The use of a revised version[9b] of the FCI with fewer ambiguities and a smaller likelihood of false positives has had little impact[17a] on $\langle g\rangle_{IE}$ as measured at Indiana and Harvard Universities. In addition, (a) interview data[9a,12a] suggest that ambiguities and false positives are relatively rare, (b) these errors would be expected to bias the IE and T courses about equally and therefore have little influence on the difference $\langle\langle g\rangle\rangle_{48IE} - \langle\langle g\rangle\rangle_{14T}$.

***2. Teaching to the Test and Test-question Leakage.***

Considering the elemental nature of the FCI questions, for IE courses both the average $\langle\langle g\rangle\rangle_{48IE} = 0.48 \pm 0.14$, and maximum $\langle g\rangle = 0.69$ are disappointingly low, and below those which might be expected if teaching to the test or test-question leakage[48] were important influences.

Of the 48 data sets[17a] for IE courses (a) 27 were supplied by respondents to our requests for data, of which 22 (81%) were accompanied by a completed survey questionnaire, (b) 13 have been discussed in the literature, and (c) 5 are Indiana University courses of which I have first-hand knowledge. All survey-form respondents indicated that they thought they had avoided "teaching to the test" in answering the question "To what extent do you think you were able to avoid 'teaching to the test(s)' (i.e., going over experiments, questions, or problems identical or nearly identical to the test items)?" Likewise, published reports of the courses in group "b" and my own knowledge of courses in group "c" suggests an absence of "teaching to the test" in the restricted sense indicated in the question. (In the broadest sense, IE courses all "teach to the test" to some extent if this means teaching so as to give students some understanding of the basic concepts of Newtonian mechanics as examined on the FCI/MD tests. However this is the bias we are attempting to measure.)

There has been no evidence of test-question leakage in the Indiana posttest results (e.g., significant mismatches for individual students between FCI scores and other course grades). So far there has been only one report[9a] of such leakage in the literature – as indicated in ref. 17a, the suspect data were excised from the survey.

***3. Fraction of Course Time Spent on Mechanics***

Comparisons can be made for T and IE courses within the same institution where the fraction $f = t_m/t_s$ of class time $t_m$ spent on mechanics (including energy and momentum conservation) to the total semester (or semester-equivalent) time $t_s$ is about the same:

Arizona State ($f = 0.8$): $\langle\langle g\rangle\rangle_{IE2} - \langle\langle g\rangle\rangle_{T3} = 0.47 - 0.24 = 0.23$;
Cal Poly ($f = 1.0$): $\langle\langle g\rangle\rangle_{IE3} - \langle\langle g\rangle\rangle_{T1} = 0.56 - 0.25 = 0.31$;
Harvard ($f = 0.6$): $\langle\langle g\rangle\rangle_{IE4} - \langle\langle g\rangle\rangle_{T1} = 0.56 - 0.27 = 0.29$;
Monroe Com. Coll. (MCC), non-calc. ($f = 0.8$): $\langle\langle g\rangle\rangle_{IE4} - \langle\langle g\rangle\rangle_{T1} = 0.55 - 0.22 = 0.33$;
MCC, calculus ($f = 1.0$): $\langle\langle g\rangle\rangle_{IE4} - \langle\langle g\rangle\rangle_{T1} = 0.47 - 0.22 = 0.25$; and
Ohio State ($f = 0.7$): $\langle\langle g\rangle\rangle_{IE1} - \langle\langle g\rangle\rangle_{T1} = 0.42 - 0.18 = 0.24$.

Thus a substantial difference $\langle\langle g\rangle\rangle_{IE} - \langle\langle g\rangle\rangle_{T}$ is maintained where the time factor is equal.

That the gain difference is not very sensitive to the fraction of the course time spent on mechanics over the range common in introductory courses can also be seen from the fact that the differences quoted above are rather similar to (a) one another despite the differences in f, and (b) the difference $\langle\langle g\rangle\rangle_{IE48} - \langle\langle g\rangle\rangle_{T14} = 0.25$ which characterizes the entire survey, despite the fact that f varies among the survey courses. Questionnaire responses covering 22 of the survey courses indicated that f ranged from 0.7 to 1.0 with an average of $0.9 \pm 0.1$sd .

***4. Post and Pretest Motivation of Students***

As indicated in "2" above, of the 48 data sets[17a] for IE courses, 27 were supplied by respondents to our requests for data, of which 22 were accompanied by a completed survey questionnaire. Responses to the question "Did the FCI posttest count as part of the final grade in your course? If so give the approximate weighting factor" were: "No" (50% of the 22 courses surveyed); "Not usually" (9%); "Yes, about 5%" (23%); "Yes, weighting factor under 10%" (9%); No Response, 9%. For the 11 courses for which *no grade incentives were offered* $\langle\langle g\rangle\rangle_{IE11} = 0.49 \pm 0.10$sd, close to the average $\langle g\rangle$ for all the 48 IE courses of the survey $\langle\langle g\rangle\rangle_{IE48} = 0.48 \pm 0.14$sd. Thus it seems doubtful that posttest grade-incentive motivation is a significant factor in determining the normalized gain.

As for the pretest, grade credit is, of course, inappropriate but $\langle g\rangle$ can be artificially raised if students are not induced[49] to take the pretest seriously. All surveyed instructors answered "Yes" to the survey form question "Do you think that your students exerted serious effort on the FCI pretest?" Likewise, published reports of the courses not surveyed and my own knowledge of courses at Indiana suggests that students did take the pretest seriously.

***5. Hawthorne/John Henry Effects***[47]

These effects can produce short-term benefits associated with (a) the special attention (rather than the intrinsic worth of the treatment) given to a research test group (Hawthorne effect), or (b) the desire of a control group to exceed the performance of a competing test group (John Henry effect). Such benefits would be expected to diminish when the treatment is applied as a regular long-term routine to large numbers of subjects. Among IE courses, Hawthorne effects should be relatively small for courses where IE methods have been employed for many years in regular instruction for hundreds of students: five 1994-5 courses at Monroe Community College[27] (N =169) ; four 1993-5 courses at Indiana University[30,34] (N = 917); and three 1993-5 courses at Harvard[29] (N = 560). For these 12 courses $\langle\langle g\rangle\rangle_{IE12} = 0.54 \pm 0.10$sd, about the same as the $\langle\langle g\rangle\rangle_{IE29} = 0.51 \pm 0.10$sd average of the 29 IE courses (excluding the 7 atypical Low-g courses) for which, on average, Hawthorne effects were more likely to have occurred. Students may well benefit from the special attention paid to them in regular IE instruction over the long term, but this benefit is intrinsic to the pedagogy and should not be classed as a Hawthorne effect. I shall not consider John Henry effects because any correction for them would only decrease $\langle\langle g\rangle\rangle_{T14}$, and thus increase the difference $\langle\langle g\rangle\rangle_{48IE} - \langle\langle g\rangle\rangle_{14T}$.

Although no reliable quantitative estimate of the influence of systematic errors seems possible under the present survey conditions, arguments in "1" to "5" above, and the general uniformity of the survey results, suggest that it is extremely unlikely that systematic error plays a significant role in the nearly two-standard-deviation difference observed in the average

normalized gains of T and IE courses shown in Eq. (2c) and in Fig. 1. *Thus we conclude that this difference primarily reflects variation in the effectiveness of the pedagogy and/or implementation.*

## VI. IMPACT OF PHYSICS-EDUCATION RESEARCH

All interactive-engagement methods used in the survey courses were stimulated in one way or another by physics- education research (PER)[51,52] and cognitive science.[44,53] It is significant that of the 12 IE courses[9a,c;21;27;29;30b,c,d;54-56] that achieved normalized gains $g \geq 0.60$ (see Figs. 1,3), 67% were taught at least in part by individuals who had devoted considerable attention to PER as judged by their publication of peer-reviewed articles or books on that subject [the same can be said for 48% of the 36 IE courses with $(\langle g \rangle) < 0.6$]. It is also noteworthy that of the 12 IE courses with $g \geq 0.60$, 42% utilized texts[27a,52b,54,55] based on PER [the same can be said for 19% of the 36 IE courses with $(\langle g \rangle) < 0.6$]. It would thus appear that PER has produced very positive results in the classroom.

For the 48 interactive-engagement courses of Fig. 1, the ranking in terms of number of IE courses using each of the more popular methods is – Collaborative Peer Instruction (CPI)[31,32]: 48 (*all* courses); Microcomputer-Based Labs (MBL)[57]: 35; Concept Tests[29]: 20; Modeling[14,58]: 19; Active Learning Problem Sets (ALPS)[28b] or Overview Case Studies (OCS)[28b]: 17; physics-education-research based text or no text: 13; and Socratic Dialogue Inducing (SDI) Labs[8,13,34]: 9. [For simplicity, courses combined[17a] into one "course" [TO (8 courses), TO-C (5 courses) and IUpre93 (5 courses) are counted as one course each.] The ranking in terms of number of students using each method is– CPI: 4458 (*all* students); MBL: 2704; Concept Tests: 2479; SDI: 1705; OCS/ALPS: 1101; Modeling: 885; research-based text or no text: 660.

A detailed breakdown of the instructional strategies as well as materials and their sources for each of the 48 IE courses of this survey is presented in a companion article.[17a] The IE methods are usually interdependent and can be melded together to enhance one another's strengths and modified to suit local conditions and preferences (especially easy if materials are available electronically[27a,29c,34c] so as to facilitate copying, pasting, and cutting). *All these IE strategies, having proven themselves to be relatively effective in large-scale pre/post testing, deserve serious consideration by physics teachers who wish to improve their courses, by physics-education researchers, and by designers of new introductory physics courses.*[58c,59]

## VII. SUGGESTIONS FOR COURSE AND SURVEY IMPROVEMENTS

Although the 48 interactive-engagement courses of Figs. 1 - 3 appear, on average, to be much more effective than traditional courses, none is in the High-g region and some are even in the Low-g region characteristic of traditional courses. This is especially disturbing considering the elemental and basic nature of the Force Concept Inventory and Mechanics Diagnostic test questions. (Many instructors refuse to place such questions on their exams, thinking that they are "too simple."[18]) As indicated above, case studies[17a] of the Low-g IE courses strongly suggest the presence of implementation problems. Similar detailed studies for Medium-g IE courses were not carried out, but personal experience with the Indiana courses and communications with most of the IE instructors in this study suggest that similar though less severe implementation problems (Sec. IIIA) were common.

Thus there appear to be no magic bullets among the IE treatments of this survey and more work seems to be required on both their content and implementation. As argued more trenchantly in ref. 17a, this survey and other work suggests that improvements may occur through, e.g., (a) use of IE methods in *all* components of a course and *tight integration* of all those components[60]; (b) careful attention to motivational factors and the provision of grade incentives for taking IE activities seriously; (c) administration of exams in which a substantial number of the questions probe the degree of conceptual understanding induced by the IE methods; (d) inexpensive augmentation of the teaching/coaching staff by undergraduate and postdoctoral students[34b,61]; (e) *apprenticeship* education of instructors new to IE methods[8,34b]; (f) early recognition and positive intervention for potential low-gain students[62]; (g) explicit focus on the goals and methods of science[2,52,63] (including an emphasis on operational definitions[2,8a,13,34,52b]); (h) more personal attention to students by means of human-mediated computer instruction in some areas[64,65]; (i) new types of courses[58c,59]; (j) advances in physics-education research and cognitive science. More generally, a *redesign process* (described by Wilson and Daviss[66] and undertaken in refs. 34 and 67) of continuous *long-term* classroom use, feedback, assessment, research analysis, and revision seems to be required for substantive educational reform.

Standards and measurement are badly needed in physics education[68] and are vital components of the redesign process. In my view, the present survey is a step in the right direction but improvements in future assessments might be achieved through (in approximate order of ease of implementation) (1) standardization of test-administration practices[48,49]; (2) use of a survey questionnaire[15c] refined and sharpened in light of the present experience; (3) more widspread use of standardized tests[9b;10;50a-c,57c] by individual instructors so as to monitor the learning of their students; (4) observation and analysis of classroom activities by independent evaluators[59a]; (5) solicitation of anonymous information from a large random sample of physics teachers; (6) development and use of new and improved versions of the FCI and MB tests, treated with the confidentiality of the MCAT,[48] (7) use of E&M concept tests,[69] and questionnaires which assess student views on science and learning[63]; and (8) reduction of possible teaching-to-the-test influence by drawing test questions from pools such that the specific questions are unknown to the instructor.[68]

## VIII. SUMMARY AND CONCLUSION

Fourteen traditional (T) courses (N = 2084) which made little or no use of interactive-engagement (IE) methods achieved an average gain $\langle\langle g \rangle\rangle_{14T} = 0.23 \pm 0.04$. In sharp contrast, forty-eight IE courses (N = 4458) which made substantial use of IE methods achieved an average gain $\langle\langle g \rangle\rangle_{48IE} = 0.48 \pm 0.14$. It is extremely unlikely that systematic errors play a significant role in the nearly two-standard-deviation difference in the normalized gains of the T and IE courses.

A plot of average course scores on the Hestenes/Wells problem-solving Mechanics Baseline test versus those on the conceptual Force Concept Inventory show a strong positive correlation with coefficient r = + 0.91. Comparison of IE and traditional courses implies that IE methods *enhance* problem-solving ability.

The conceptual and problem-solving test results strongly suggest that *the use of IE strategies can increase mechanics-course effectiveness well beyond that obtained with traditional methods.*

### Epilogue

This survey indicates that the strenuous recent efforts to reform introductory physics instruction, enlightened by cognitive science and research in physics education, have shown very positive results in the classroom. However, history[70-75] suggests the possibility that such efforts may have little lasting impact. This would be most unfortunate, considering the current imperative to (a) educate more effective science majors[76] and science-trained professionals,[77] and (b) raise the appallingly low level of science literacy[78,79] among the general population. Progress towards these goals should increase our chances of solving the monumental science-intensive problems[80-86] (economic, social, political, and environmental) that beset us, but major upgrading of physics education on a national scale will probably require (1) the cooperation of instructors, departments, institutions, and professional organizations,[87] (2) long-term classroom use, feedback, assessment, research analysis, and redesign of interactive-engagement methods.[66]


## ACKNOWLEDGMENTS

This work received partial support from NSF Grant DUE/MDR9253965. My deepest gratitude goes to those teachers who supplied the invaluable advice, manuscript suggestions, and unpublished data which made this report possible (see ref. 17a for details of their work): Albert Altman, Dewayne Beery, Les Bland, Don Boys, Ben Brabson, Bernadette Clemens-Walatka, Paul D'Alessandris, Randall Knight, Priscilla Laws, Cherie Lehman, Eric Mazur, Roger Mills, Robert Morse, Piet Molenaar, Tom O'Kuma, Gregg Swackhamer, Lou Turner, Alan Van Heuvelen, Rick Van Kooten, Mojtaba Vazari, William Warren, and Paul Zitzewitz. I have benefited from additional suggestions by Amit Bhattacharyya, Ernie Behringer, Sister Marie Cooper, Steve Gottlieb, Ibrahim Halloun, John Hardie, David Hammer, Charles Hanna, David Hestenes, Don Lichtenberg, Tim Long, Joe Redish, Rudy Sirochman, Steve Spicklemire, Richard Swartz, Jack Uretsky, and Ray Wakeland. I thank two discerning AJP referees for constructive criticism which considerably improved the manuscript. This work would never have been completed without the encouragement and counsel of Arnold Arons, David Hestenes, William Kelly, and Ray Hannapel.

a) Electronic mail: <hake@ix.netcom.com>

1. (a) I. Halloun and D. Hestenes, "The initial knowledge state of college physics students," Am. J. Phys. **53**, 1043-1055 (1985); corrections to the Mechanics Diagnostic test are given in ref. 14; (b) "Common sense concepts about motion," *ibid.* **53**, 1056-1065 (1985).

2. A.B. Arons, *A Guide To Introductory Physics Teaching* (Wiley, 1990); reprinted with minor updates in *Teaching Introductory Physics* (Wiley, 1997). The latter book also contains *Homework and Test Questions for Introductory Physics Teaching* (Wiley, 1994) along with a new monograph "Introduction to Classical Conservation Laws."

3. F. Reif, "Educational Challenges for the University," Science **184**, 537-542 (1974); "Scientific approaches to science education," Phys. Today **39**(11), 48-54 (1986).

4. D. Hestenes, "Wherefore a Science of Teaching," Phys. Teach. **17**, 235-242 (1979).

5. J. Clement, "Students' preconceptions in introductory mechanics," Am. J. Phys. **50**, 66-71 (1982).

6. M. McClosky, "Intuitive Physics," Sci. Am. **248**(4), 122-130 (1983).

7. L.C. McDermott, "Research on conceptual understanding in mechanics," Phys. Today **37**(7), 24-32 (1984).

8. R.R. Hake, Indiana University, (a) "Promoting student crossover to the Newtonian world," Am J. Phys. **55**, 878-884 (1987); (b) "My Conversion To The Arons-Advocated Method Of Science Education," Teaching Education **3**(2), 109-111 (1991).

9. (a) D. Hestenes, M. Wells, and G. Swackhamer, Arizona State University, "Force Concept Inventory," Phys. Teach. **30**, 141-158 (1992). The FCI is very similar to the earlier Mechanics Diagnostic test and pre/post results using the former are very similar to those using the latter. (b) I. Halloun, R.R. Hake, E.P. Mosca, and D. Hestenes, Force Concept Inventory (Revised, 1995) in ref. 29b. (c) Gregg Swackhamer, Glenbrook North High School (public), private communication, 4/96.

10. D. Hestenes and M. Wells, "A Mechanics Baseline Test," Phys. Teach. **30**, 159-166 (1992).

11. (a) D. Huffman and P. Heller, "What Does the Force Concept Inventory Actually Measure?" Phys. Teach. **33**, 138-143 (1995); (b) P. Heller and D. Huffman, "Interpreting the Force Concept Inventory: A Reply to Hestenes and Halloun," *ibid.* **33**, 503, 507-511 (1995).

12. (a) D. Hestenes and I. Halloun, "Interpreting the Force Concept Inventory: A Response to March 1995 Critique by Huffman and Heller," Phys. Teach. **33**, 502, 504-506 (1995); (b) I. Halloun and D. Hestenes, "The Search for Conceptual Coherence in FCI data," preprint, 1996.

13. S. Tobias and R.R. Hake, Indiana University, "Professors as physics students: What can they teach us?" Am. J. Phys. **56**, 786-794 (1988).

14. I. A. Halloun and D. Hestenes, Arizona State University, "Modeling instruction in mechanics," Am. J. Phys. **55**, 455-462 (1987). The ASU-HH-C point of Fig. 3c is for "Test Group #3."

15. R.R. Hake, (a) "Assessment of Introductory Mechanics Instruction," AAPT Announcer **23**(4), 40 (1994); (b) "Survey of Test Data for Introductory Mechanics Courses," *ibid.* **24**(2), 55 (1994); (c) "Mechanics Test Data Survey Form," 15 pages, copies available on request. The form's list of physics-education strategies and resources may be useful.

16. PHYS-L and PhysLrnR are networks of respectively, physics teachers and physics-education researchers. Alan Cairns gives instructions on subscribing to these and other such e-mail discussion groups at <http://www-hpcc.astro.washington.edu/scied/physics/physlists.html>.

17. R.R. Hake, (a) "Interactive engagement methods in introductory mechanics courses," preprint, 5/97, available on request; (b) "Evaluating conceptual gains in mechanics: A six-thousand student survey of test data," *Proceedings of the 1996 International Conference on Undergraduate Physics Education* (College Park, MD, in press); in that paper the 7 Low-g IE courses, deemed to have implementation problems as evidenced by instructors' comments, were omitted from the IE averaging so as to obtain $\langle\langle g \rangle\rangle_{41IE} = 0.52 \pm 0.10$sd. I now think that the present treatment is preferable.

18. E. Mazur, "Qualitative vs. Quantitative Thinking: Are We Teaching the Right Thing?" Optics and Photonics News **3**, 38 (1992).

19. B.S. Bloom, "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," Educational Researcher **13**(6), 4-16 (1984): "Using the standard deviation (sigma) of the control (conventional) class, it was typically found that the average student under tutoring was about two standard deviations above the average of the control class....The tutoring process demonstrates that *most* of the students do have the potential to reach this high level of learning. I believe an important task of research and instruction is to seek ways of accomplishing this under more practical and realistic conditions than the one-to-one tutoring, which is too costly for most societies to bear on a large scale. This is the '2 sigma' problem."

20. A.M. Saperstein, Phys. Teach. **33**, 26-27 (1995) has analyzed data in ref. 9a to conclude that FCI pretest scores increase by about 3 - 4% per year in the late teen years simply due to life experience (not formal instruction).

21. Bernadette Clemens-Walatka, Sycamore High School (public), private communication, 3/96.

22. Lou Turner, Western Reserve Academy, private communication, 3/96. Western Reserve Academy is selective private coeducational prep school.

23. Priscilla Laws, Dickinson College, "Calculus-Based Physics Without Lectures," Phys. Today **44**(12), 24-31 (1991); "Millikan Lecture 1996: Promoting active learning based on physics education research in introductory physics courses," Am. J. Phys. **65**, 13-21 (1997).

24. Don Boys, Univ. of Michigan (Flint), private communication, 8/94; M. Vaziri and D. Boys "Improving the Introductory Mechanics Course," AAPT Announcer **24**(2), 81 (1994).

25. D.E. Meltzer and K. Manivannan, "Promoting Interactivity in Physics Lecture Classes," Phys. Teach. **34**, 72-76 (1996).

26. E. Cullota, "The Calculus of Educational Reform," Science **255**, 1060-1062 (1992);

27. Paul D'Alessandris, Monroe Community College (MCC), (a) private communications 11/94, 5/95, 3/96, 4/96; used his own physics-education-research based workbooks (available at <FTP@eckert.acadcomp.monroecc.edu>) in place of a course text. No grade incentives for performance on the posttest FCI were given at MCC. (b) P. D'Alessandris, "The Development of Conceptual Understanding and Problem-Solving Skills through Multiple Representations and Goal-less Problems," AAPT Announcer **24**(4), 47 (1994). MCC is an open admissions two-year college which "draws....urban poor, suburban underachievers, and rural everything else."

28. Alan Van Heuvelen, calculus-based course at Ohio State University, (a) private communication, 4/96. No grade incentives for performance on the posttest FCI were given at Ohio State. (b) A. Van Heuvelen, "Overview, Case Study Physics," Am. J. Phys. **59**, 898-907 (1991). (c) "Experiment Problems for Mechanics," Phys. Teach. **33**, 176-180 (1995).

29. Eric Mazur, Harvard University, calculus-based course for science (but not physics) majors, (a) private communications, 5/95, 4/96; (b) *Peer Instruction: A User's Manual* (Prentice Hall, 1997), contains the 1995 Revised FCI. (c) For assessment data, course syllabus, information on Classtalk, and examples of Concept Tests see <http://galileo.harvard.edu/>.

30. Unpublished data for non-calculus based Indiana University courses for science (but not physics) majors, enrolling primarily pre-meds and pre-health-professionals: (a) IU93S, R.R.Hake, Spring 1993; (b) IU94S R.R. Hake, Spring 1994; used the physics- education-research based text and workbook by Reif (ref. 52b); (c) IU95S: R. Van Kooten, R.R. Hake, F.M. Lurie, and L.C. Bland, Spring 1995; (d) IU95F: L.C. Bland, B.B. Brabson, R.R.Hake, J.G. Hardie, and E. Goff, Fall 1995. At Indiana, the FCI posttest normally counts for half the final-exam grade (about 12% of the final-course grade).

31. D.W. Johnson, R.T. Johnson, and K.A. Smith, *Cooperative Learning: Increasing College Faculty Instructional Productivity* (George Washington University, 1991).

32. P. Heller, R. Keith, S. Anderson, "Teaching problem solving through cooperative grouping, Part 1: Group vs individual problem solving," Am. J. Phys. **60**, 627-636 (1992); P. Heller and M. Hollabaugh "Teaching problem solving through cooperative grouping, Part 2: Designing problems and structuring groups," *ibid*., p. 637-644.

33. Classtalk. a classroom communication system used at Harvard and other institutions, is discussed by J.C. Webb, G.R. Webb, R Caton, and F. Hartline, "Collaborative Learning and Insights on Students' Thinking: Computer Technology Offers New Possibilities for Large Classes," AAPT Announcer **24**(4), 64 (1994).

34. R.R. Hake (a) "Socratic Pedagogy in the Introductory Physics Lab," Phys. Teach. **30**, 546-552 (1992); (b) "Socratic Dialogue Labs in Introductory Physics," in *Proceedings of the 1995 Conference on New Trends in Physics Teaching*, ed. by J. Slisko (Univ. of Puebla; Puebla, Mexico, in press). (c) For a summary of recent work and an updated listing of electronically available SDI materials (e.g., manuals, teacher's guides, sample lab exams, equipment set-up lists) see <http://carini.physics.indiana.edu/SDI/> or contact R. R. Hake at <hake@ix.netcom.com>. Ref. 34a and SDI Labs #1-3 (versions of 10/93) are available on the Fuller-Zollman CD-ROM InfoMall. (d) A. Bhattacharyya, R.R. Hake, R. Sirochman, "Improving Socratic Dialogue Inducing (SDI) Labs," AAPT Announcer **25**(2), 80 (1995). (e) A Grading Acronym Guide sheet is available on request.

35. (a) J.L. Uretsky, "Using 'Dialogue Labs' in a Community-College Physics Course," Phys. Teach. **31**, 478-481 (1993); (b) N. C. Steph, "Improving the Instructional Laboratory with TST and SDI Labs: Mixing, Matching, and Modifying Ideas," AAPT Announcer **21**(4), 61 (1991). (TST ≡ Tools for Scientific Thinking.)

36. Concept Test implementation at Indiana University is discussed in ref. 17a.

37. A. Roychoudhury, D. Gabel, and R.R. Hake, "Inducing and Measuring Conceptual Change in Introductory-Course Physics Students," AAPT Announcer **19**(4), 64 (1989).

38. At Indiana, both cooperative group problem-solving in recitations and the "Physics Forum" were initiated by Professor H.O. Ogren.

39. First Class, in extensive use for large-enrollment classes at Indiana University, allows electronic-bulletin-board discussions, file sharing, and collaboration among students and instructors.

40. C. Schwartz, unpublished work describing this physicist's invention of "Minute Papers" as described by R.C. Wilson, "Improving Faculty Teaching," J. of Higher Ed. **57**(2), 196-211 (1986) and private communication. For a discussion see ref. 17a .

41. R.R. Hake and J. C. Swihart, "*DI*agnostic *S*tudent *CO*mputerized *E*valuation of Multicomponent Courses (DISCOE)" Teaching and Learning (Indiana University), January 1979, available on request.

42. Figure 4 shows two points which are atypical in that they lie close to the diagonal with an average MB score only a few points below the average FCI score: (1) ASU-VH105-C for Van Heuvelen's Physics 105 course for disadvantaged pre-engineering students at Arizona State Univ., (2) UML94-C for Albert Altman's calculus-based course at UMass (Lowell). These points can be explained in terms of unusual course features: Van Heuvelen's methods are oriented towards problem solving, while Altman's classes contain a relatively high percentage (about 30%) of non-native English speakers whose scores on the FCI may be artificially lowered by language difficulties. That the other UML93-C point is close to the least-squares fit line can be accounted for by the fact that in 1993 (unlike in 1994) *no grade credit was given* for performance on the MB exam, even though the MB exam requires the "intolerable labor of thought" to a much greater extent than the FCI. (Grade credit *was* given at UML in 1993 and 1994 for performance on the FCI posttest.)

43. B. Thacker, E. Kim, K. Trefz, and S.M. Lea, "Comparing problem solving performance of physics students in inquiry-based and traditional introductory physics courses," Am. J. Phys. **62**, 627-633 (1994).

44. E.F. Redish, "Implications of cognitive studies for teaching physics," Am. J. Phys. **62**, 796-803 (1994).

45. See e.g., (a) J.R. Taylor, *Introduction to Error Analysis* (University Science Books, 1982), esp. p. 87-89, 126-130; (b) L. Kirkup, *Experimental Methods* (Wiley, 1994), esp. p. 85 -87. These conventional methods are not strictly applicable because the pre- and posttest score distributions tend to depart from Gaussians– our experience is that FCI posttest scores are usually negatively skewed (ceiling effect).

46. For simplicity let $\langle S_f \rangle \equiv x$ and $\langle S_i \rangle \equiv y$, so that Eq. (1) becomes $\langle g \rangle = (x - y)/(C - y)$ where C is the number of questions on the exam and x and y are the average number of correct responses on the post and pretests. The conventional treatment (e.g., ref. 45) regards the random error $\Delta x$ to be the "standard deviation of the mean." This is just the $sd_x$ divided by the square root of the number of measurements N. Thus $\Delta x = sd_x / N^{1/2}$ and $\Delta y = sd_y / N^{1/2}$. If one thinks of measuring $x_a$, $x_b$, $x_c$, ..... $x_n$ for many sets n (say $10^6$) of N students (all drawn randomly from the same enormous hypothetical homogeneous population) and then taking the average of $x_a$, $x_b$, $x_c$, ..... $x_n$ to define the "true" value x(true) of x, then one can be 68% confident that $x_a$ lies within $\pm\Delta x$ of x(true). With this definition of $\Delta x$ (and similarly for $\Delta y$), the conventional treatment then specifies $\Delta\langle g \rangle = \{[(\partial\langle g \rangle/\partial x)\Delta x)]^2 + [(\partial\langle g \rangle/\partial y)\Delta y)]^2\}^{1/2}$. Here $\partial\langle g \rangle/\partial x = 1/(C - y)$ and $\partial\langle g \rangle/\partial y = (x - C)/(C - y)^2$ . All this assumes that the separate errors associated with x and y are random and independent, and are small enough that they may be treated as differentials. For the presently analyzed data, the average, minimum, and maximum of: $\Delta x$ = 2.0%, 0.6%, and 3.5% of C; $\Delta y$ = 2.0%, 0.4%, and 3.8% of C.

As suggested by R.E. Mills, somewhat lower random errors are entailed if one takes the average g for a course to be $g(ave) \equiv (1/N) \Sigma_j g_j = (1/N) \Sigma_j (post_j - pre_j)/(C - pre_j)$. In practice, for $N \geq 20$, g(ave) is usually within 5% of $\langle g \rangle$. Eq. (1) and the above definition of g(ave) imply that $[g(ave) - \langle g \rangle]$ is proportional to the $g_j$ -weighted average of the deviations $(pre_j - \langle pre_j \rangle)$. Since the average of $(pre_j - \langle pre_j \rangle)$ is zero, a low $[g(ave) - \langle g \rangle]$ implies a low correlation between $g_j$ and $pre_j \equiv (S_i)_j$ for individual students, just as there is a low correlation between $\langle g \rangle$ and $\langle S_i \rangle$ for courses, as discussed just above Eq. (2c).

47. See, e.g., R.E. Slavin, *Research Methods in Education* (Allyn and Bacon, 2nd ed., 1992).

48. It is unfortunate that the national-assessment value of arduously constructed and validated standardized tests such as the FCI and the MB is gradually being eroded by distribution of answers to students at some institutions. The danger of question leakage is especially severe if the posttest FCI/MB scores are used to determine part of the final course grade. At Indiana, the FCI test is always given and referred to as a "diagnostic mechanics exam" in an attempt to shield ref. 9a. We collect *all* pre-and posttests from students and none is returned. The pre- and post-tests scores are posted by ID, but questions and answers are neither posted, disseminated, nor shown as computer animations. After the posttest, instructors are quite willing to discuss FCI/MB questions privately with any student, but answer keys are not posted. Because there are many sources (ref. 17a) of good conceptual questions, there is little need to draw on the standardized tests for questions to be used for ordinary class discussion and testing. Indiana students understand that the FCI must be treated just as the MCAT, and there is little dissatisfaction. Because of the above mentioned dispersal of answers at some institutions, and the fact that the FCI and MB tests were published in the open literature, their useful lives may not extend for more than another year. New and better tests (treated with the confidentially of the MCAT) are sorely needed in time for a calibration against the original or revised FCI. The necessary steps in the laborious process of constructing valid and reliable multiple-choice physics tests have been discussed in refs. 1a, 9a, and 50.

49. At Arizona (ref. 1a) and Indiana (ref. 8) it is explained to students that their scores on the pretest will not count towards the course grade but will be confidentially returned to them and will assist both themselves and their instructors to know the degree and type of effort required for them to understand mechanics.

50. (a) R.J. Beichner, "Testing student interpretation of kinematics graphs," Am. J. Phys. **62**, 750-762 (1994); (b) S.J. Sobolewski, "Development of Multiple-Choice Test Items," Phys. Teach. **34**, 80-82 (1996); (c) W. Pfeiffenberger, A.M. Zolandz, and L. Jones, "Testing Physics Achievement: Trends over Time and Place," Phys. Today **44**(9), 30-37 (1991); (d) G.J. Aubrecht, "Is There a Connection Between Testing and Teaching?" J. Coll. Sci. Teach. **20**, 152-157 (1991); G.J. Aubrecht and J.D. Aubrecht, "Constructing Objective Tests," Am. J. Phys. **51**, 613-620 (1983).

51. For overviews of physics-education research see, e.g., (a) *Toward a scientific practice of science education*, ed. by M. Gardner, J.G. Greeno, F. Reif, A.H. Schoenfeld, A. diSessa, and E. Stage (Erlbaum, 1990); (b) *Research in Physics Learning: Theoretical Issues and Empirical Studies*, R. Duit, F. Goldberg, and H. Niedderer, eds. (Institute for Science Ed., Kiel, 1992); (c) A. Van Heuvelen "Learning to think like a physicist: A review of research-based instructional strategies," Am. J. Phys. **59**, 891-897 (1991); (d) A. B. Arons, "Generalizations to be drawn from results of research on teaching and learning" in *Thinking Physics for Teaching*, ed. by C. Bernardini, C. Tarsitani, and M. Vicintini (Plenum, 1995); (e) D. Hammer, "More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research," Am. J. Phys. **64**, 1316-1325 (1996).

52. F. Reif, (a) "Millikan Lecture 1994: Understanding and teaching important scientific thought processes," Am. J. Phys. **63**, 17-32 (1995); (b) *Understanding Basic Mechanics* (Text, Workbook, and Instructor's Manual) (Wiley, 1994).

53. J. Mestre and J. Touger, "Cognitive Research--What's in It for Physics Teachers?" Phys. Teach. **27**, 447-456 (1989).

54. Randall Knight, California Polytechnic State University (San Luis Obispo), private communications 4/94, 3/96; used his own physics-education-research based text *Physics: A Contemporary Perspective* (Addison-Wesley, 1997).

55. A.L. Ellermeijer, B. Landheer, P.P.M. Molenaar, "Teaching Mechanics through Interactive Video and a Microcomputer-Based Laboratory," 1992 NATO Amsterdam Conference on Computers in Education, Springer Verlag, in press; private communications from P.P.M. Molenaar, 6/94, 4/96. Used the physics-education-research based J.A. Dekker, *Motion* (in Dutch) (Univ. of Amsterdam, 1990).

56. Thomas O'Kuma, Lee College, (a) private communication, 5/95, 4/96. Lee is an open admissions 2-year college with a majority of students from low to low middle income families. It has over 30% minorities, over 56% women students, an average student age of 29, and (according to O'Kuma) is fairly typical of most two-year community colleges.

57. (a) R.F. Tinker, "Computer Based Tools: Rhyme and Reason," in *Proc. Conf. Computers in Physics Instruction*, ed. by E. Redish and J. Risley (Addison-Wesley, Reading, MA, 1989), pp. 159-168; (b) R.K. Thornton and D. R. Sokoloff, "Learning motion concepts using real-time microcomputer-based laboratory tools," Am. J. Phys. **58**, 858-867 (1990); (c) D.R. Sokoloff, P.W. Laws, and R.K. Thornton, "Real Time Physics, A New Interactive Introductory Lab Program," AAPT Announcer **25**(4), 37 (1995).

58. (a) I. A. Halloun and D. Hestenes, "Modeling instruction in mechanics," Am. J. Phys. **55**, 455-462 (1987); (b) D. Hestenes, "Toward a modeling theory of physics instruction," *ibid*. **55**, 440-454 (1987); "Modeling Games in the Newtonian World," *ibid.* **60**, 732-748 (1992). (c) M. Wells, D. Hestenes, and G. Swackhamer, "A modeling method for high school physics instruction, *ibid.* **63**, 606-619 (1995), <http://modeling.la.asu.edu/modeling.html>.

59. See, e.g., (a) J.S. Rigden, D.F. Holcomb, and R. DiStefano, "The Introductory University Physics Project," Phys. Today **46**(4), 32-37 (1993). R. DiStefano, "The IUPP Evaluation: What we were trying to learn and how we were trying to learn it," Am. J. Phys. **64,** 49-57 (1996); "Preliminary IUPP results: Student reactions to in-class demonstrations and to presentations of coherent themes," *ibid.*, **64,** 58-68 (1996). (b) R.P. Olenick, "C3P (Comprehensive Conceptual Curriculum for Physics)," AAPT Announcer **26**(2), 68 (1996), <http://phys.udallas.edu>. Other citations appear in ref. 17a.

60. R.R. Hake, "Towards Mastery of Newtonian Mechanics by the Average Student," AAPT Announcer **24**(1), 23 (1994).

61. I have found top-notch undergraduate physics majors, *after suitable apprenticeships*, to be among the best IE instructors, evidently because their minds are closer to those of the students and they have only recently struggled to understand introductory physics concepts themselves. Thus they can better appreciate the nature and magnitude of the intellectual hurdles and ways to overcome them. Undergraduates have the further advantage that they are relatively inexpensive to employ. Post-doctoral students have also volunteered to serve as lab instructors, since they are often motivated to seek experience with advanced educational methods in order to better qualify themselves for job opportunities in the expanding market for educationally-effective teachers. As future professionals, the undergraduate, graduate, and post-doctoral student instructors all provide the opportunity to seed interactive-engagement methods into science education at all levels.

62. R.R. Hake, R. Wakeland, A. Bhattacharyya, and R. Sirochman, "Assessment of Individual Student Performance in an Introductory Mechanics Course," AAPT Announcer **24**(4), 76 (1994). Scatter plots of gains (posttest - pretest) vs pretest scores for all students in a class delineate relatively high-g (low-g) students for whom the course was (was not) effective. We discuss various diagnostic tests (mechanics, mathematics, and spatial visualization) given to incoming students which might be used to recognize *potential* "low gainers" and thus initiate helpful intervention.

63. I. Halloun, "Views About Science and Physics Achievement: The VASS Story," *Proceedings of the 1996 International Conference on Undergraduate Physics Education* (College Park, MD, in press); E.F. Redish, R.N. Steinberg, and J.M. Saul, "The Distribution and Change of Student Expectations in Introductory Physics," *ibid.*; I. Halloun and D. Hestenes, "Interpreting VASS Dimensions and Profiles," Sci. and Ed., in press.

64. See, e.g., (a) A. A. diSessa, "The Third Revolution in Computers and Education," J. Res. in Sci. Teach. **24**, 343-367 (1987); (b) J.J. Kaput, "Technology and Mathematics Education" in *Handbook of Research on Mathematics Teaching and Learning*, D.A. Grouws, ed. (MacMillan, 1992); (c) R.D. Pea, "Augmenting the Discourse of Learning with Computer-Based Learning Environments," in *Computer-Based Learning Environments and Problem Solving*, ed. by E. DeCorte, M.C. Linn, H. Mandl, and L. Verschaffel (NATO ASI Series, series F, vol. 84); (d) E. Redish and J. Risley, eds., *Proc. Conf. Computers in Physics Instruction* (Addison-Wesley, 1989).

65. (a) R. Bird and R.R. Hake, "Force Motion Vector Animations on the Power Mac,"AAPT Announcer **25**(2), 80 (1995); (b) R.R. Hake and R. Bird, "Why Doesn't The Water Fall Out Of The Bucket? Concept Construction Through Experiment, Discussion, Drawing, Dialogue, Writing, and Animations," *ibid.* **25**(2), 70 (1995).

66. K.G. Wilson and B. Daviss, *Redesigning Education* (Henry Holt, 1994); a goldmine of valuable references. See also at <http://www-physics.mps.ohio-state.edu/~kgw/RE.html>.

67. L. C. McDermott, (a) "Millikan Lecture 1990: What we teach and what is learned: Closing the gap," Am. J. Phys. **59**, 301-315 (1991); (b) L.C. McDermott, P.S. Shaffer, and M.D. Somers, "Research as a guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine," *ibid*., **62**, 46-55 (1994).

68. F. Reif, "Guest Comment: Standards and measurements in physics – Why not in physics education?" Am. J. Phys. **64**, 687-688 (1996).

69. (a) C.J. Hieggelke, D. Maloney, T. O'Kuma, and A. Van Heuvelen, "Electric and Magnetic Concept Inventory," AAPT Announcer **26**(2), 80 (1996); (b) P.V. Engelhardt and R.J. Beichner, "Determining and Interpreting Students' Concepts of Resistive Electric Circuits," *ibid.* **26**(2), 80-81 (1996).

70. A. B. Arons, "Uses of the Past: Reflections on United States Physics Curriculum Development, 1955 to 1990," Interchange **24**(1&2), 105-128 (1993); "Improvement of Physics Teaching in the Heyday of the 1960's," in *Conference on the Introductory Physics Course on the occasion of the retirement of Robert Resnick,* Jack Wilson, ed. (Wiley, 1997), p. 13-20.

71. L. C. McDermott, "A perspective on teacher preparation in physics and other sciences: The need for special science courses for teachers," Am. J. Phys. **58**, 734-742 (1990).

72. C. Swartz, "The Physicists Intervene," Phys. Today **44**(9), 22-28 (1991): "For over 150 years American physicists have been making forays into elementary and high school science teaching. Their novel approaches have usually worked--*but the results have always been short-lived.*" (Our *italics*.)

73. S. Tobias, "Guest Comment: Science Education Reform: What's wrong with the process?" Am. J. Phys. **60**, 679-681 (1992); *Revitalizing Undergraduate Science: Why Some Things Work and Most Don't* (Research Corporation, 1992).

74. S.B. Sarason, *The Predictable Failure of Educational Reform* (Jossey-Bass, 1990); *Revisiting "The Culture of The School and The Problem of Change"* (Teachers College Press, 1996).

75. G. Holton, "A Nation at Risk, Revisited" in *The Advancement of Science and its Burdens* (Univ. of Cambridge Press, 1986).

76. W.P. Wolf, "Is Physics Education Adapting to a Changing World?" Phys. Today **47**(10), 48-55 (1994).

77. S. Tobias, " 'Science-Trained Professionals' – A New Breed for the New Century," J. Sci. Ed. Technol. **5**, 167-169 (1996).

78. "Science literacy," as used here, does *not* mean the knowledge of science "facts" as measured by some "science literacy tests," but rather an understanding of the methods, history, and limitations of science; the relationship of science to society and to other disciplines; and a working knowledge of science in at least a few areas such as to allow further self-education as the need may arise. See Arons, ref. 2, p. 289-290, for a more complete discussion.

79. G. Holton, "The Anti-Science Phenomenon," (and citations therein) in *Science and Anti-Science* (Harvard University Press, 1993); *Einstein, History, and Other Passions: The Rebellion Against Science at the End of the Twentieth Century* (Addison Wesley, 1996).

80. R. Marshall and M. Tucker, *Thinking for a Living* (Basic Books, 1992).

81. A.A. Bartlett, "Reflections on Sustainability, Population Growth, and the Environment," Population and Environment **16**(1), 5-34 (1994).

82. M. Gell-Mann, *The Quark and the Jaguar: Adventures in the Simple and the Complex* (W.H. Freeman, 1994), ch. 22, pp. 345 - 366.

83. G.E. Brown, "New Ways of Looking at US Science and Technology," Phys. Today **47**(9), 31-35 (1994).

84. R.W. Schmitt, "Public Support of Science: Searching for Harmony," Phys. Today **47**(1), 29-33 (1994).

85. "Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology," Advisory Committee to the NSF Directorate for Education and Human Services, 1996, available at <http://www.ehr.nsf.gov/EHR/DUE/documents/review/96139/start.htm> or as a hard copy by request to <undergrad@NSF.gov>.

86. "Preparing for the 21st Century: The Education Imperative," National Research Council, 1997, available at <http://www2.nas.edu/21st>.

87. R.C. Hilborn, "Physics at the Crossroads: Innovation and Revitalization in Undergraduate Physics – Plans for Action," report on a College Park AAPT conference of 9/96; "Guest Comment: Revitalizing undergraduate physics – Who needs it?" Am. J. Phys. **65**, 175-177 (1997).